

**THE UNIVERSITY OF DANANG
UNIVERSITY OF SCIENCE AND TECHNOLOGY
FACULTY OF ADVANCED SCIENCE AND TECHNOLOGY**



ĐỒ ÁN TỐT NGHIỆP - CAPSTONE PROJECT PERSONALIZED TEXT-TO-IMAGE GENERATION

PHAM THI THUY DUONG – CLASS: 20PFIEV2

Advanced Undergraduate Program in Industrial Informatics under the
Programme de Formation d'Ingénieurs d'Excellence au Vietnam (PFIEV)

Supervisor: PhD. Nguyen Quang Nhu Quynh
Co-Supervisors: Prof. Wen-Nung Lie
(National Chung Cheng University)

Submitted to the Faculty of Advanced Science and Technology in Partial
Fulfillment of the Requirements for the Degree of Engineer

Danang, 06/ 2025

**THE UNIVERSITY OF DANANG
UNIVERSITY OF SCIENCE AND TECHNOLOGY
FACULTY OF ADVANCED SCIENCE AND TECHNOLOGY**

ĐỒ ÁN TỐT NGHIỆP - CAPSTONE PROJECT

PERSONALIZED TEXT-TO-IMAGE GENERATION

PHAM THI THUY DUONG – CLASS: 20PFIEV2

Advanced Undergraduate Program in Industrial Informatics under the
Programme de Formation d'Ingénieurs d'Excellence au Vietnam (PFIEV)

Supervisor: PhD. Nguyen Quang Nhu Quynh
Co-Supervisors: Prof. Wen-Nung Lie
(National Chung Cheng University)

Danang, 06/ 2025

Số: 882 /QĐ-ĐHBK

Đà Nẵng, ngày 03 tháng 3 năm 2025

QUYẾT ĐỊNH

**Về việc thành lập Hội đồng hướng dẫn Capstone project cho sinh viên
ngành Kỹ thuật điện - chuyên ngành Tin học công nghiệp,
Khoa Khoa học Công nghệ tiên tiến**

HIỆU TRƯỞNG TRƯỜNG ĐẠI HỌC BÁCH KHOA

Căn cứ Nghị định số 32/CP ngày 04 tháng 4 năm 1994 của Chính phủ về việc thành lập Đại học Đà Nẵng;

Căn cứ Thông tư số 10/2020/TT-BGDĐT ngày 14 tháng 5 năm 2020 của Bộ trưởng Bộ Giáo dục và Đào tạo ban hành Quy chế tổ chức và hoạt động của đại học vùng và các cơ sở giáo dục đại học thành viên;

Căn cứ Nghị quyết số 08/NQ-HĐĐH ngày 12 tháng 7 năm 2021 của Hội đồng Đại học Đà Nẵng về việc ban hành Quy chế tổ chức và hoạt động của Đại học Đà Nẵng và Nghị quyết 13/NQ-HĐĐH ngày 07 tháng 9 năm 2021 của Hội đồng Đại học Đà Nẵng về việc sửa đổi, bổ sung một số điều của Quy chế tổ chức và hoạt động của Đại học Đà Nẵng;

Căn cứ Nghị quyết số 93/NQ-HĐT ngày 18 tháng 12 năm 2023 của Hội đồng trường Trường Đại học Bách khoa ban hành Quy chế tổ chức và hoạt động của Trường Đại học Bách khoa, Đại học Đà Nẵng;

Căn cứ Quyết định số 840/QĐ-ĐHBK ngày 28 tháng 02 năm 2023 của Hiệu trưởng Trường Đại học Bách khoa, Đại học Đà Nẵng ban hành Quy định về việc tổ chức và triển khai thực hiện Thực tập tốt nghiệp và Đồ án tốt nghiệp theo hình thức Capstone Project của Trường Đại học Bách khoa, Đại học Đà Nẵng;

Căn cứ Biên bản cuộc họp ngày 21 tháng 02 năm 2025 về việc Họp xét giao Đồ án tốt nghiệp học kỳ 2, năm học 2024-2025;

Căn cứ Tờ trình số 29/TTr-KHCNTT ngày 28 tháng 02 năm 2025 của Trường khoa Khoa Khoa học Công nghệ tiên tiến về việc giao đề tài Capstone project cho sinh viên ngành Kỹ thuật điện - chuyên ngành Tin học công nghiệp, Khoa Khoa học Công nghệ tiên tiến;

Theo đề nghị của Trường phòng Phòng Đào tạo.

QUYẾT ĐỊNH:

Điều 1. Thành lập Hội đồng hướng dẫn Capstone project cho 14 sinh viên khóa 2020, Chương trình đào tạo kỹ sư chất lượng cao Việt - Pháp, ngành Kỹ thuật điện - chuyên ngành Tin học công nghiệp, Khoa Khoa học Công nghệ tiên tiến. Tên đề tài, sinh viên thực hiện và danh sách Hội đồng hướng dẫn được đính kèm theo Quyết định.

Điều 2. Hội đồng hướng dẫn tổ chức cho sinh viên thực hiện Capstone project theo quy định hiện hành. Sau khi hoàn thành nhiệm vụ, Hội đồng tự giải thể.

Điều 3. Trường phòng Phòng Tổ chức - Hành chính, Trường phòng Phòng Đào tạo, Trường phòng Phòng Kế hoạch - Tài chính, Trường khoa Khoa Khoa học Công nghệ tiên tiến và viên chức, sinh viên có tên ở Điều 1 căn cứ Quyết định thi hành./.

Nơi nhận:

- Như Điều 3;
- Khoa Khoa học Công nghệ tiên tiến;
- Lưu: VT, ĐT.

KT. HIỆU TRƯỞNG

PHÓ HIỆU TRƯỞNG



PGS.TS. Nguyễn Hồng Hải

ĐẠI HỌC ĐÀ NẴNG
TRƯỜNG ĐẠI HỌC BÁCH KHOA

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập – Tự do – Hạnh phúc

Số: 882 /QĐ-ĐHBK

Đà Nẵng, ngày 03 tháng 3 năm 2025

QUYẾT ĐỊNH

Về việc thành lập Hội đồng hướng dẫn Capstone project cho sinh viên
ngành Kỹ thuật điện - chuyên ngành Tin học công nghiệp,
Khoa Khoa học Công nghệ tiên tiến

ĐỀ TÀI 3			
1. Tên đề tài: Personalized text-to-image generation (Tạo hình ảnh từ văn bản được cá nhân hóa). (Đề tài thực hiện và bảo vệ bằng Tiếng Anh)			
2. Sinh viên thực hiện	Phạm Thị Thùy Dương	20PFIEV2	
3. Hội đồng hướng dẫn	TS. Nguyễn Quang Như Quỳnh	Giảng viên, Khoa Khoa học Công nghệ tiên tiến, Trường Đại học Bách khoa	Chủ tịch – Hướng dẫn
	GS Wen-Nung Lie	Đại học Quốc lập Trung Chính, Đài Loan (NATIONAL CHUNG CHENG UNIVERSITY, TAIWAN)	Đồng hướng dẫn

P F I E V

GRADUATION PROJECT - EVALUATION FORM
PHIẾU CHẤM ĐỒ ÁN TỐT NGHIỆP CAPSTONE PROJECT

(Form used for Supervision Committee to evaluate each student upon completion of the graduation project/

General information

Student's name: Student's name: Pham Thi Thuy Duong; Class: 20PFIEV2

- Project's title: Personalized Text-to-Image Generation

Comments for the graduation project:

1. Importance, novelty, applicability of the thesis

Personalized text-to-image generation is a highly promising topic in the field of artificial intelligence and computer vision, with broad applications such as creation, entertainment, fashion, medical imaging, and identity reconstruction. The author proposed a new method called Fusion Cross Initialization (Fusion-CI)

2. Achievement of the required tasks and objectives of the thesis:

The author proposed a new method called Fusion Cross Initialization (Fusion-CI). and achieved experiment result which is consistently achieved higher identity similarity compared to Cross Initialization, particularly for prompts involving facial expressions—a challenging aspect of identity preservation.

3. Form, structure, layout of the thesis:

Well-written, logical structure

4. Scientific value/ publication/ addressing the problems posed by the enterprise or the university:

This thesis was developed in 2025 at the Department of Electrical Engineering, National Chung Cheng University (CCU), under the TEEP Internship Program. Fusion-CI offers a meaningful improvement in terms of maintaining personal identity, which is essential in applications where consistent appearance across varied prompts is crucial which will be continued developed in next phase in CCU

5. Weakness and points to improve:

III. Student attitude: Good in learning attitude, creativity

IV. Evaluation of the Supervisor Committee:

1. The grading given by the Chairman of the Supervision Committee: 10

The grading given by the co-supervisor: 9.3

The average grade of the Supervision Committee: 9.7/10 (take up to 1 decimal number)

2. Recommendation: ☒ Approved for thesis defense ☐ Not approved for thesis defense

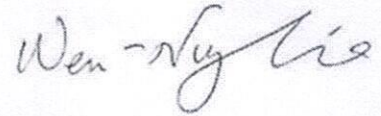
- Further comments:

Da Nang, date 19 month 06 year 2025

Full name &
signature of the
Chairman of the
Supervision
Committee


Nguyễn Đức Nhứt

Full name & signature of the
co-supervisor



PHIẾU CHẤM PHẢN BIỆN ĐỒ ÁN TỐT NGHIỆP CAPSTONE PROJECT

(Mẫu dùng cho UV phản biện của Hội đồng bảo vệ ĐATN)

I. Thông tin chung:

- Họ tên sinh viên: Phạm Thị Thùy Dương, Lớp: 20PFIEV2
- Tên đề tài: Personalized Text-to-Image Generation

II. Nhận xét đồ án tốt nghiệp:

1. Về tính cấp thiết, tính mới, khả năng ứng dụng của đề tài:

The thesis addresses the urgent and novel problem of text-to-image generation with a focus on personalization to preserve individual facial features. This is highly relevant as current models (e.g., DALL·E, Imagen, Stable Diffusion) struggle to maintain identity consistency. The approach has wide applicability in e-commerce, entertainment, education, character design, and personal media protection.

2. Về kết quả giải quyết các nội dung nhiệm vụ yêu cầu của đề án:

The thesis successfully accomplishes its objectives:

- Comprehensive analysis of diffusion models, Textual Inversion, and Cross Initialization techniques.
- Proposal of the "Fusion Cross Init" method combining textual and visual information via a Transformer module.
- Successful implementation on top of Stable Diffusion, with both qualitative and quantitative experiments (improved CLIP scores and ArcFace similarity).
- A functional demo system was developed.

3. Về hình thức, cấu trúc, bố cục của đồ án tốt nghiệp:

The thesis is clearly structured and logically organized, including background, theoretical foundation, proposed methods, experiments, and evaluation. Illustrations and diagrams (e.g., on pages 28, 40, 47) are clear and professional. The formatting adheres to academic standards, with complete appendices, lists of figures, tables, and references.

4. Đề tài có giá trị khoa học/ có bài báo/ giải quyết vấn đề đặt ra của doanh nghiệp hoặc nhà trường:

The work has high scientific value, proposing a new embedding initialization strategy that enhances accuracy and personalization in generated images. It has potential for publication in international conferences on AI, computer vision, or multimodal learning. The project also addresses practical problems relevant to industry and academia.

5. Các tồn tại, thiếu sót cần bổ sung, chỉnh sửa:

Should add analysis on computational cost and runtime performance, which are crucial for real-world deployment. A broader evaluation on more subjects and diverse prompts is needed to confirm generalizability.

III. Đánh giá của người phản biện:

Mẫu 9 - Mẫu chấm điểm của người phản biện (cuối kỳ, M3)

1. Điểm chấm: 9.7/10 (lấy đến 1 số lẻ thập phân)

2. Đề nghị: ☒ Được bảo vệ đồ án ☐ Không được bảo vệ

IV. Câu hỏi đề nghị sinh viên trả lời trong buổi bảo vệ:

1. Why did you choose to base your method on Stable Diffusion rather than other diffusion models?
2. Could you describe in detail how the Fusion Transformer module improves semantic-visual alignment?
3. How did you select the evaluation metrics (CLIP score and ArcFace similarity), and why are they suitable for this task?

Đà Nẵng, ngày 26 tháng 6 năm 2025

Họ tên & chữ ký người phản biện



Nguyễn Thị Anh Thư

ABSTRACT

Text-to-Image Generation is a promising application of artificial intelligence, enabling the automatic creation of images from natural language descriptions. A key direction in this field is personalized generation, where a textual prompt and a reference image of a target subject (typically a face) are used to synthesize identity-preserving images that semantically match the input. However, current models like DALL·E [1], Imagen [2], and Stable Diffusion [3] struggle to preserve facial identity - especially with complex prompts - often producing distorted or inaccurate faces.

To address this, prior methods such as Textual Inversion [4] rely on random initialization, resulting in weak identity retention. Cross Initialization [5] improves this by using semantic information from the subject’s name, but remains limited to known identities within the training data. To overcome these limitations, I propose an improved token initialization strategy that integrates linguistic information with visual features extracted from a reference image. Specifically, I introduce `fusion_cross_init`-a novel method that combines both textual and visual information at the token initialization stage. This method incorporates a Fusion Transformer module to enhance the alignment between the visual features of the reference image and the semantic information related to the subject. Furthermore, it leverages CLIP’s [6] multi-modal encoders to synchronize semantic and visual signals from the outset, thereby significantly improving identity consistency in personalized text-to-image generation. The method is implemented on top of the pre-trained Stable Diffusion [3] model, ensuring high image quality while optimizing computational efficiency.

The proposed system is evaluated through a combination of qualitative analysis and two key quantitative metrics: CLIP Score [60], which measures the semantic alignment between the prompt and the generated image, and ArcFace Similarity [7], which evaluates the degree of identity preservation between the reference and generated images. The overall results indicate that my approach significantly improves both facial identity retention and prompt-image consistency compared to existing methods. Future work will aim to improve generalization to diverse subjects and complex prompts.

ACKNOWLEDGMENT

This thesis was developed in 2025 at the Department of Electrical Engineering, National Chung Cheng University, under the TEEP Internship Program. The research was conducted under the supervision of Prof. Wen-Nung Lie (賴文能) and PhD. Nguyen Quang Nhu Quynh.

I would like to express my sincere gratitude to Prof. Wen-Nung Lie and PhD. Quynh for their invaluable guidance, insightful feedback, and continuous support throughout the course of this project. Their expertise and encouragement have been essential in shaping the direction and success of my research. I am also deeply thankful to senior student Liao Tze-Lin (廖子駢) and the Chung Cheng University lab members for their generous support and guidance, which helped me adapt to the new environment and acquire new knowledge during my internship.

Special thanks to PhD. Nguyen Le Hoa, Dean of the Faculty, and to PhD. Le Quoc Huy, along with all the teachers who taught me at the University of Science and Technology – The University of Danang, for their dedicated teaching, which provided me with a strong academic foundation for this research. I am also grateful to PhD. Nguyen Thi Anh Thu, Assoc. Prof. PhD. Pham Van Tuan, and L.Y.D.I.N.C Ltd. Company for offering me early research opportunities, which greatly contributed to my readiness for this thesis.

I would also like to express my appreciation to Prof. Lie's laboratory for providing computing equipment and a supportive research environment. I am thankful to National Chung Cheng University for hosting me and providing the infrastructure needed for this project. I gratefully acknowledge the Taiwan Ministry of Education for awarding me the TEEP scholarship, which financially supported my stay and research activities during this internship.

Finally, I would like to thank my family for their unconditional support, which allowed me to focus fully on my research. Special thanks also go to my friends and colleagues, whose insightful discussions and encouragement turned this research journey into a meaningful and memorable experience.

This thesis would not have been possible without the dedication, support, and contributions of all the individuals and institutions mentioned above.

Danang, June 2025

The University of Science and Technology – The University of Danang

Pham Thi Thuy Duong



DECLARATION

I hereby declare that the entirety of this graduation thesis is the result of my own independent research and effort, carried out seriously and under the guidance of my academic supervisor.

All data, information, and results presented in this thesis have been collected, analyzed, and presented with honesty, accuracy, and objectivity. I affirm that no part of this work has been copied, plagiarized, or used unlawfully from any other sources. All references and materials consulted in the process of conducting this research have been properly cited and acknowledged in accordance with academic integrity standards.

I understand that any violation of academic ethics, including plagiarism or data fabrication, will result in disciplinary action as prescribed by the university's regulations on scientific research and academic honesty.

I take full responsibility for the content of this thesis.

Danang, June 2025

The University of Science and Technology – The University of Danang

Pham Thi Thuy Duong

Table of Contents

ABSTRACT	i
ACKNOWLEDGMENT.....	ii
DECLARATION.....	iv
List of Figures	vii
List of Tables	viii
Abbreviations.....	ix
INTRODUCTION.....	x
1. Motivation.....	x
2. Contribution	xi
3. Roadmap	xii
4. Outline	xiv
CHAPTER 1	1
LITERATURE REVIEW.....	1
1.1 Text-to-Image Synthesis.....	1
1.2 Image Inversion in Text-to-Image Personalization	2
1.3 Personalization in Text-to-Image Diffusion Models	3
1.4 Proposed Solution	4
CHAPTER 2	6
FOUNDATIONS AND TECHNICAL APPROACH TO PERSONALIZED TEXT-TO-IMAGE GENERATION	6
2.1 Preliminaries	7
2.1.1. Latent Diffusion Models (LDMs).....	7
2.1.1.1 Theoretical Foundation: From DDPM to LDM.....	8
2.1.1.2 Transition to Latent Space: Latent Diffusion Models (LDMs)	9
2.1.2 Text Embeddings	14
2.1.3 Textual Inversion method	16
2.1.3.1 Technical Details of Textual Inversion.....	17
2.1.3.2 Strengths and Extensions	18
2.1.4 The Cross Initialization method.....	19
2.1.4.1 Core Mechanism of Cross Initialization	20
2.1.4.2 Impact and Outcomes	24
2.2 Proposed method.....	24
2.2.1 Analysis.....	24
2.2.1.1 Motivation for the Proposed Method.....	24

2.2.1.2 Theoretical Foundation	25
2.2.2 Enhanced Token Initialization via Text-Image Fusion (Fusion Cross Init)	27
2.2.2.1 Input Data:	29
2.2.2.2 Detailed Structure:	29
2.2.2.3 Illustration of Workflow	40
CHAPTER 3	41
EXPERIMENTAL RESULTS & EVALUATION	41
3.1 Experiment.....	41
3.1.1 Implementation Details and Training Configuration.....	42
3.1.2 Evaluation Setup	43
3.1.2.1 Prompt Similarity Metric – CLIP Score	44
3.1.2.2 Identity Similarity – ArcFace-Based Evaluation	46
3.2 Results.....	49
3.2.1 Qualitative Evaluation	50
3.2.2 Quantitative evaluation.....	51
3.2.3. Runtime & Computational Efficiency	55
3.3 Further Experiments with Different Inputs.....	57
CONCLUSIONS	61
BIBLIOGRAPHY	64

P F I E V

List of Figures

Figure 1.	An example of the operation process of the LDM	10
Figure 2.	Architecture of Latent Diffusion Model with Cross-Attention Conditioning	12
Figure 3.	Outline of the text-embedding and inversion process	17
Figure 4.	Comparison of Textual Inversion Initialization and Cross Initialization techniques.	20
Figure 5.	Cross Initialization Method structure	21
Figure 6.	Fusion Cross Initialization Method structure	28
Figure 7.	Flowchart of FUSION CROSS INIT operation process simulation	40
Figure 8.	Example of an evaluation matrix model: CLIP-I, CLIP-T	45
Figure 9.	Input image: Person 1	49
Figure 10.	Input image: Person 2	50
Figure 11.	Extended Evaluation Results on Additional Test Cases of the Fusion Cross Initialization Method	56

List of Tables

Table 1.	Implementation plan of the project	xiii
Table 2.	Experimental Results for Person 1: Cross Initialization Method & Fusion Cross Initialization Method	49
Table 3.	Experimental Results for Person 2 : Cross Initialization Method & Fusion Cross Initialization Method	50
Table 4.	Comparison of Evaluation Metrics Between Two Methods: Cross Initialization Method & Fusion Cross Initialization Method	52
Table 5.	Extended Comparison of Evaluation Metrics Between Cross Initialization and Fusion Cross Initialization Methods	54



Abbreviations

CI	Cross Initialization
Fusion CI / Fusion Cross Init	Fusion Cross Initialization
LDM	Latent Diffusion Model
CLIP	Contrastive Language–Image Pretraining
FFN	Feed-Forward Network
BOS	Beginning of Sequence
EOS	End of Sequence
PAD	Padding
BLIP	Bootstrapped Language-Image Pretraining
EDT	Encoder-based Domain Tuning
MHSA	Multi-Head Self-Attention
RTX	Ray Tracing Texel eXtreme (NVIDIA GPU branding, not a technical acronym)
DALL·E	Not an acronym; coined from “Dalí” (artist) and “WALL·E” (robot character)
DDPM	Denoising Diffusion Probabilistic Model
CLIP-T	CLIP Text (used for text evaluation directions)

INTRODUCTION

1. Motivation

Amid the rapid advancement of artificial intelligence (AI), the ability to automatically and personally generate content has become increasingly important across various sectors, including e-commerce, media, education, gaming, and healthcare. In particular, Text-to-Image Generation has emerged as one of the most promising AI technologies, enabling the transformation of natural language descriptions into high-quality images [3], [8].

According to a 2023 report by MarketsandMarkets, the global generative AI market is projected to grow from USD 20.9 billion in 2024 to USD 136.7 billion by 2030, at a compound annual growth rate (CAGR) of 36.7% [9]. Grand View Research estimates that the global AI image generation market was valued at USD 349.6 million in 2023 and is expected to grow at a CAGR of 17.7% from 2024 to 2030 [10]. Data Insights Market projects that this market will reach USD 2 billion by 2025 and USD 20 billion by 2033, growing at a CAGR of approximately 35% between 2025 and 2033 [11]. Furthermore, a report from Global Market Estimates [12] forecasts that the text-to-image generation market will grow at a CAGR of 17.4% from 2023 to 2028.

According to XenonStack's 2023 report, Text-to-Image Generation is being widely adopted in e-commerce, advertising, and product design, enabling highly personalized user experiences [13]. Vogue Business reported in 2023 that the fashion industry is leveraging this technology to design and prototype garments, enhancing creative efficiency [14]. As published in *Nature Scientific Reports* (2024), models like Stable Diffusion and DALL·E support the visualization of scientific content—particularly in physics and biology—making learning more engaging and accessible [15], [3]. A 2024 report by SkimAI highlights the use of Text-to-Image Generation in gaming and entertainment for the rapid creation of environments, characters, and props, enhancing creativity and reducing production time [16]. These figures highlight the rapidly growing demand for Text-to-Image Generation, with expanding

applications across industries such as e-commerce, fashion, education, entertainment, and healthcare.

Despite significant progress, advanced models like DALL·E [1], Imagen [2], and Stable Diffusion [3] still face limitations in personalization-particularly in accurately preserving the facial features of a target subject and aligning with the semantic content of the input text [4], [5]. These shortcomings hinder the models' usability in domains requiring high precision, such as character design, image restoration, and identity recognition.

Therefore, developing Text-to-Image Generation systems that can effectively personalize outputs while ensuring semantic consistency with the input description is essential to meet the increasing demand for customized visual content.

2. Contribution

Personalized Text-to-Image Generation is a key area in generative AI, especially for applications that demand both identity preservation and semantic consistency [17], [8]. In this project, I propose an enhanced embedding initialization method that combines linguistic cues (e.g., name-based embeddings [5]) with visual features extracted from a single reference image. A Fusion Transformer module is introduced to better align the subject's semantic and visual representations. Additionally, the method leverages CLIP's multimodal encoders [51] to synchronize textual and visual signals from the beginning, significantly improving identity consistency in the generation process. The approach is implemented on top of the pre-trained Stable Diffusion model [3], ensuring high image quality while maintaining computational efficiency. It effectively addresses the limitations of prior methods such as Textual Inversion [4] and Cross Initialization [5], offering notable improvements in both identity fidelity and semantic alignment.

The specific contributions of this project are as follows:

- Proposed a multimodal initialization strategy that combines name-based semantic embeddings with visual features from a reference image to construct more accurate and stable personalized concept tokens.

- Introduced a Fusion Transformer module to effectively align semantic information from text with visual cues, enhancing the quality of the initial embedding.
- Implemented the proposed system on a pre-trained Stable Diffusion model to maintain high image quality while minimizing computational cost.
- The proposed initialization method is evaluated through both qualitative analysis and quantitative metrics, including CLIP Score and ArcFace Similarity. CLIP Score measures how well the generated image aligns with the input prompt, while ArcFace Similarity evaluates how effectively the subject's facial features are preserved from the reference image. Experimental results show that the proposed method significantly improves identity preservation while maintaining high consistency with the textual description.

At the time of writing, this work has produced the following outputs:

- An ongoing thesis titled “*Personalized Text-to-Image Generation*”, conducted under the TEEP Internship Program at National Chung Cheng University.
- A working demo system capable of generating identity-consistent images from natural language descriptions and a single reference image.

3. Roadmap

The implementation of this thesis was divided into four main tasks:

1. Research on the Stable Diffusion model and analysis of existing personalized *Text-to-Image Generation* techniques, including *Textual Inversion*, *DreamBooth*, and *Cross Initialization*.
2. Design and development of an enhanced embedding initialization method that integrates name-based semantic information and visual features from a single reference image.

3. Developed a Fusion Transformer module to enhance the alignment between semantic information from text and visual features from the reference image, thereby improving the quality of the initial embedding.
4. Evaluation of the proposed initialization method on Stable Diffusion, combining qualitative analysis and quantitative metrics (CLIP Score and ArcFace Similarity) to assess semantic alignment and identity preservation.

The execution timeline of each activity is presented in Table 1. All tasks and contributions were independently completed by the author, as detailed below:

Table 1. Implementation plan of the project

Task	Task Description	Timeline
1.	Research on the Stable Diffusion model architecture and analysis of existing personalized text-to-image generation techniques, including Textual Inversion, DreamBooth, and Cross Initialization.	February – March 2025
2.	Designed and developed an enhanced embedding initialization method that integrates name-based semantic information and visual features from a single reference image.	March – May 2025
3.	Developed a Fusion Transformer module to enhance the alignment between semantic information from text and visual features from the reference image, thereby improving the quality of the initial embedding.	April – June 2025
4.	Evaluation of the proposed initialization method on Stable Diffusion, combining qualitative analysis and quantitative metrics (CLIP Score and ArcFace Similarity) to assess semantic alignment and identity preservation.	May – June 2025

All technical components and experimental designs were independently executed by the author. The specific tasks and contributions are summarized as follows:

- Task 1: Literature Review and Problem Analysis: Investigated the architecture and working mechanism of the Stable Diffusion model, with a focus on its text-to-image generation pipeline. Studied personalization methods such as Textual Inversion, Cross Initialization, and DreamBooth. Identified their limitations in terms of facial identity preservation and semantic alignment. Defined research objectives and directions for improvement.
- Task 2: Design and Implementation of the Enhanced Embedding Initialization Method: Developed a multimodal initialization strategy that combines name-based semantic embeddings (generated using a fine-tuned CLIP text encoder) with visual features extracted from a single reference image. Integrated the resulting personalized token into Stable Diffusion through cross-attention mechanisms and embedding injection (hooks).
- Task 3: Development of Fusion Transformer Module: Developed a Fusion Transformer module to enhance the alignment between semantic information from text and visual features extracted from the reference image. This module improves the quality of the initial embedding, leading to better identity consistency and semantic representation in the generated images.
- Task 4: Qualitative and Quantitative Evaluation: Evaluated the proposed embedding initialization method on a pre-trained Stable Diffusion model through qualitative and quantitative analysis. The quantitative evaluation used two metrics: CLIP Score to assess semantic alignment between the prompt and the generated image, and ArcFace Similarity to measure identity preservation relative to the reference image. Compared with the Cross Initialization method, the proposed approach showed notable improvements in both facial identity fidelity and semantic consistency.

4. Outline

This thesis is organized into three chapters. Chapter 1 introduces the research background and motivation for personalized Text-to-Image Generation, highlights the main contributions of the thesis, and presents the project roadmap. Chapter 2

provides a comprehensive overview of related work and the proposed methodology, including existing approaches such as Textual Inversion and Cross Initialization, an overview of the Stable Diffusion framework, as well as the enhanced embedding initialization technique, the multimodal integration process, and the development of the Fusion Transformer Module. Finally, Chapter 3 describes the experimental setup, evaluation metrics, and a comparison of results with baseline methods.



CHAPTER 1

LITERATURE REVIEW

Currently, various approaches have been developed for the task of text-to-image generation, including GAN-based models, autoregressive transformers, diffusion models, and retrieval-based methods. These systems leverage large-scale training datasets and deep generative architectures to synthesize images from textual descriptions. Among them, diffusion-based models such as DALL·E 2 [1], Imagen [2], and Stable Diffusion [3] have demonstrated superior performance in generating high-quality images with strong text-image alignment.

However, existing methods still face significant challenges, particularly in personalized generation tasks where the goal is to faithfully preserve the identity of a specific individual. A lack of control over identity-related attributes, such as facial features, often leads to distorted or semantically inconsistent outputs [18, 19]. Moreover, personalization techniques often require extensive fine-tuning or embedding learning from limited reference data, resulting in low efficiency and a high susceptibility to overfitting [18].

To address these limitations, recent approaches have focused on improving identity preservation and enhancing semantic alignment. These methods typically leverage available semantic priors and integrate textual, visual, and pretrained model representations [18, 19]. This chapter provides an overview of prominent personalization techniques, initialization strategies, and multimodal embedding methods that are currently being explored to enhance the performance of personalized text-to-image generation.

1.1 Text-to-Image Synthesis

Text-to-image (T2I) generation is a rapidly evolving field within generative AI, aiming to synthesize realistic images conditioned on natural language descriptions. This task requires deep semantic understanding of the input text and the ability to generate visually coherent content that aligns with the described scenes,

objects, and relationships [20, 21]. Over the years, various model architectures have been proposed, each contributing significant advancements to image realism, resolution, and text-image alignment.

Early approaches were predominantly based on Generative Adversarial Networks (GANs), such as StackGAN [22] and AttnGAN [23], which utilized text embeddings to guide image generation through a two-stage pipeline. While GAN-based methods offered good fidelity, they struggled to capture complex semantics and maintain global coherence [24].

Later, autoregressive transformer models like DALL·E [20] introduced more robust capabilities by modeling the joint distribution of text and image tokens using large-scale datasets. These models demonstrated stronger compositionality but often suffered from slow inference speeds.

More recently, diffusion models-notably DALL·E 2 [1], Imagen [2], and Stable Diffusion [3]-have set new state-of-the-art performance benchmarks in high-fidelity image generation. Diffusion-based approaches operate by gradually denoising a random noise vector conditioned on text input. Their ability to generate diverse, realistic images with strong semantic alignment has made them the dominant paradigm in current T2I research [20, 25].

Alongside these, retrieval-based methods have emerged, leveraging large image-text datasets to retrieve and assemble relevant visual content rather than generating it from scratch. While faster, these methods often lack flexibility in generating novel compositions [26].

1.2 Image Inversion in Text-to-Image Personalization

Image inversion is a fundamental technique in generative modeling, particularly for tasks involving image manipulation and personalization. It refers to the process of mapping a real image back into the latent space of a pretrained generator such that the generator can reconstruct the original image from this latent code. This enables downstream tasks like editing, attribute control, and, in this context, personalized image synthesis [27, 28, 29].

In the context of GANs and other generative models, inversion can be accomplished in two primary ways: (1) by directly optimizing a latent code to minimize the reconstruction error between the generated and original images [27, 28, 29], or (2) by using encoder networks that learn to map images into the latent space [30, 31, 32, 33, 34, 35, 36]. These techniques have been widely applied in various image editing tasks [28, 37, 38].

With the rise of diffusion models, inversion techniques have adapted to new requirements. In diffusion-based frameworks, image inversion involves estimating an initial noise latent that, when passed through the denoising process, reproduces the original image [20, 39, 40]. This noise latent then serves as the foundation for guided manipulation, such as text-based editing and concept personalization [41, 42, 43, 44, 45].

1.3 Personalization in Text-to-Image Diffusion Models

Personalization in text-to-image diffusion models refers to the adaptation of pre-trained generative models to synthesize images of specific, unseen subjects based on a few reference images. This capability is essential for generating images that preserve individual identity while allowing for flexible manipulation based on new text prompts [4, 18].

Common techniques for personalization include:

- Textual embedding optimization, such as Textual Inversion [4], which learns a new token embedding that captures the identity of a subject from a few images. It does so without modifying the model weights, instead introducing pseudo-tokens representing new concepts.
- Model fine-tuning approaches, like DreamBooth [18] or Custom-Edit [46], which adapt parts of the diffusion model-typically the attention layers or U-Net-to incorporate novel subject concepts via subject-specific prompt tuning.
- Encoder-based methods, such as Encoder-based Domain Tuning (EDT) [47], which aim to efficiently map new image concepts to text embeddings using

pretrained encoders, allowing faster personalization compared to full model retraining.

These methods enable a wide range of applications including image editing [48], stylization, and even personalized 3D content generation [49]. However, many approaches suffer from overfitting, especially in low-shot settings where only a few reference images are available. This typically results in limited generalization to novel prompts and poor semantic alignment between the personalized embedding and the text description [50].

To address these challenges, Cross Initialization was proposed by Pang et al. [5]. Rather than initializing the token embedding randomly (as in Textual Inversion), this method starts with a semantically meaningful embedding derived from the CLIP text encoder [6]—typically initialized with general or specific category names such as "a person" or even the name of a celebrity. This strategy narrows the gap between the initial and learned embeddings, thereby stabilizing training and improving both identity preservation and prompt alignment. Additionally, the method includes regularization terms to ensure consistency in response to varying prompts and support more effective editability.

However, a notable limitation of Cross Initialization is that its success depends heavily on the semantic proximity between the initialization text and the target identity. When applied to novel subjects without strong semantic priors (e.g., non-celebrities, abstract concepts), the linguistic initialization is insufficient, leading to degraded performance in terms of both identity fidelity and prompt responsiveness [5, 50].

1.4 Proposed Solution

To address the limitations of existing personalization methods—such as Textual Inversion, which relies on randomly initialized embeddings, and Cross Initialization, which depends solely on linguistic priors—this study proposes an enhanced multimodal initialization strategy. Instead of relying exclusively on textual input, my

method integrates both semantic text features and visual cues extracted from reference images to construct a more robust and identity-aware embedding.

This approach is implemented on top of a pre-trained Stable Diffusion model [37], leveraging its strong generative capabilities while avoiding expensive retraining. The process begins with a language-based embedding generated via the CLIP text encoder [6], initialized using subject-specific information (e.g., a name or category label). This embedding is then refined using visual features obtained from the CLIP image encoder, forming a multimodal concept token that balances linguistic intent and visual identity—a design inspired by BLIP [5] and OpenCLIP [6].

To further improve the alignment between semantic intent and visual identity, I developed a Fusion Transformer module that effectively merges textual embeddings with visual features extracted from the reference image. This module enhances the coherence between modalities, enabling the system to generate concept tokens that better capture both semantic meaning and subject-specific visual characteristics. The resulting embedding improves the model's ability to maintain identity fidelity while adapting to varied and complex prompts.

In summary, my method introduces a principled fusion of linguistic and visual modalities during initialization, resulting in faster convergence, reduced overfitting, and more faithful identity preservation in personalized text-to-image generation. I evaluated its effectiveness on a pre-trained Stable Diffusion model using two key quantitative metrics: CLIP Score to assess semantic alignment between the prompt and the generated image, and ArcFace Similarity to measure the degree of identity preservation relative to the reference image.

CHAPTER 2

FOUNDATIONS AND TECHNICAL APPROACH TO PERSONALIZED TEXT-TO-IMAGE GENERATION

Embedding initialization plays a vital role in personalized text-to-image generation, as it directly impacts the model's ability to preserve visual identity and accurately interpret the semantic content of the input prompt. Achieving both objectives simultaneously is a significant challenge, especially when only one or a few reference images are available. Therefore, a well-designed initialization strategy is essential to ensure both identity consistency and semantic fidelity in the generated outputs.

This chapter first introduces the theoretical foundations of personalized image synthesis using latent diffusion models. It then provides a critical analysis of existing embedding initialization methods, highlighting both their strengths and limitations in terms of identity preservation and prompt alignment. These techniques, which often rely on random initialization or solely on linguistic information, tend to fall short in accurately capturing the subject's visual features.

To address these limitations, this chapter presents the proposed method—Fusion Cross Initialization (Fusion-CI) which combines visual features extracted from a reference image with semantic features encoded by the CLIP text encoder. This multimodal fusion approach is designed to produce a more balanced and expressive initialization embedding, thereby enhancing the personalization capabilities of the diffusion model.

By the end of this chapter, readers will gain a comprehensive understanding of the research motivation, theoretical framework, and technical formulation behind the proposed embedding initialization method, setting the foundation for the experimental evaluation in the subsequent chapters.

2.1 Preliminaries

To support the proposed personalization strategy, this section introduces the foundational concepts and components upon which I build my method. First, I provide an overview of Latent Diffusion Models (LDMs), which serve as the backbone of modern text-to-image generation systems [52], [3]. I then explain how text embeddings are used to condition image generation based on natural language prompts [20]. Next, I review the Textual Inversion technique, a commonly used baseline for learning personalized embeddings from a limited number of reference images [4]. Following that, I present Cross Initialization, a more recent approach that incorporates semantic priors (e.g., names or textual descriptions) into the initialization process [5]. Together, these components lay the groundwork for my proposed multimodal initialization method, which aims to enhance facial identity preservation and semantic alignment in personalized image synthesis.

2.1.1. Latent Diffusion Models (LDMs)

In this project, I build upon the Stable Diffusion model [3], which represents a significant advancement in the field of text-to-image generation. Stable Diffusion is based on Latent Diffusion Models (LDMs), which are optimized versions of Denoising Diffusion Probabilistic Models (DDPMs) [51]. These models have achieved impressive results in generating high-resolution, semantically aligned images from natural language prompts.

The Stable Diffusion model synthesizes images through a denoising process that gradually transforms pure noise into realistic images [3]. This is achieved by training a neural network to reverse a fixed forward diffusion process in which Gaussian noise is incrementally added to an image over multiple steps. To reduce the computational complexity of traditional pixel-space diffusion models, Stable Diffusion operates in a compressed latent space produced by a powerful autoencoder [3]. This latent space retains essential semantic content while discarding high-frequency noise, making training and inference more efficient.

2.1.1.1 Theoretical Foundation: From DDPM to LDM

To understand Stable Diffusion, it is important to grasp the underlying theory of Denoising Diffusion Probabilistic Models (DDPMs) [51] and how Latent Diffusion Models (LDMs) extend them. This section walks through the key equations and intuition.

Stable Diffusion is built on denoising diffusion probabilistic models (DDPMs), a class of generative models that learn to reverse a gradual noising process [51]. In the forward diffusion process, clean data x_0 is progressively corrupted by Gaussian noise. Formally, one defines a fixed variance schedule $0 = \beta_0 < \beta_1 < \dots < \beta_T$. At each step t , the noised image x_t is sampled from:

$$q(x_t | x_{t-1}) = N(x_t; \sqrt{1 - \beta_t} \cdot x_{t-1}, \beta_t \cdot I) \quad (2.1)$$

With:

x_0 : the original clean image

x_t : image (or feature vector) after noise is added at step t

x_{t-1} : image from previous step

β_t : noise variance at step, usually lies in the small interval $[0, 0.02]$

$N(\mu, \Sigma)$: Normal distribution with mean μ and covariance Σ

I : identity covariance matrix

After many steps $x_T \approx N(0, I)$ (pure noise) [51]. Equivalently, one can show that x_t can be directly sampled from the original image x_0 via:

$$q(x_t | x_0) = N(x_t; \sqrt{1 - \alpha_t} \cdot x_0, (1 - \alpha_t) \cdot I) \quad (2.2)$$

where $\alpha_t = \prod_{i=1}^t 1 - \beta_i$

In contrast, the reverse diffusion (generation) process is a learned Markov chain that gradually denoises from x_T back to x_0 . This reverse chain is parameterized by a neural network ϵ_θ that predicts, at each step t , a Gaussian distribution:

$$p_\theta(x_{t-1} | x_t) = N(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \quad (2.3)$$

With the mean and covariance output by the model [51], one can parameterize the reverse process. In practice, one often fixes the variance and trains the model to predict the added noise ϵ_θ in the forward step. For example, Ho et al. (2020) showed that a model ϵ_θ , trained to match the true noise ϵ , minimizes the variational bound and yields high-quality samples [1]. In summary, the DDPM framework learns a denoising mapping from noised data x_t to less-noisy x_{t-1} , enabling generation by starting from pure noise and iteratively removing it.

$$L_{\text{simple}}(\theta) = E_{x_0, \epsilon, t} [\| \epsilon - \epsilon_\theta(x_t, t) \|^2] \quad (2.4)$$

With:

- ϵ : Gaussian noise sampled during the forward process
- $\epsilon_\theta(x_t, t)$: noise predicted by the model
- t : the timestep index

2.1.1.2 Transition to Latent Space: Latent Diffusion Models (LDMs)

A major challenge with DDPMs is the high computational cost of training and inference in pixel space. Latent Diffusion Models address this by learning and performing the diffusion process in a compressed latent space z , which captures the perceptually relevant features of the image while removing superfluous details. [3]

- Autoencoder (ϵ, D): Maps an image x into latent representation $z = \epsilon(x)$, and reconstructs via $\hat{x} = D(z)$
- Latent diffusion: The diffusion process then occurs on z rather than on raw pixels, making the process far more efficient [3].

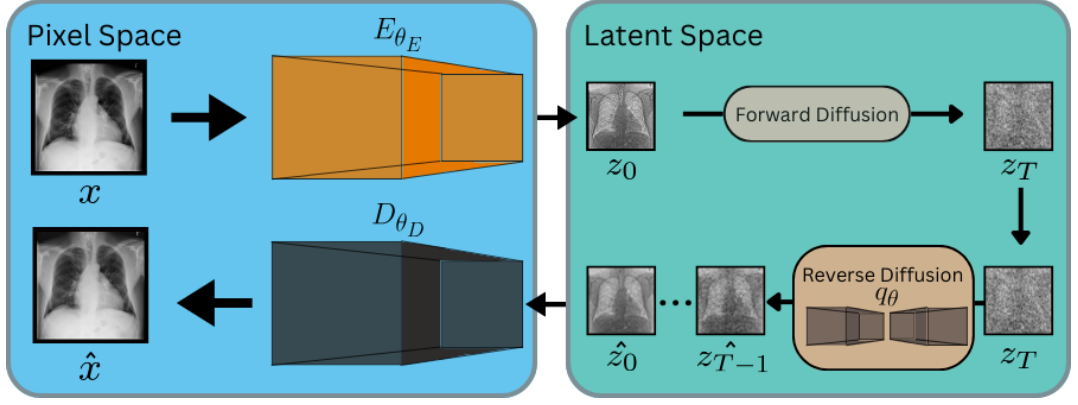


Figure 1. An example of the operation process of the LDM (OpenAI).

In my case, I utilize a pre-trained autoencoder provided by the Stable Diffusion framework. [3]

(1) Training Objective.

In the conditional Stable Diffusion setting (e.g., text-to-image), the Latent Diffusion Model (LDM) is trained to predict the noise added to a latent representation of the input image. Let y be a conditioning input (such as a text prompt), and let $\tau_{\theta}(y)$ denote its learned embedding produced by a text encoder. For a training sample x , its latent representation is $z_0 = \varepsilon(x)$, where ε is the encoder.

At each training step, a random timestep $t \in \{1, \dots, T\}$ is chosen, and Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$ is sampled. The noisy latent z_t is computed as:

$$z_t = \sqrt{\alpha_t} \cdot z_0 + \sqrt{1 - \alpha_t} \cdot \epsilon \quad (2.5)$$

The UNet denoiser ϵ_{θ} then takes $(z_t, t, \tau_{\theta}(y))$ as input and predicts the noise. The LDM loss is the mean-squared error between the true noise and the model's prediction [3]:

$$L_{LDM} = \mathbb{E}_{z \sim \varepsilon(x), y, \epsilon \sim \mathcal{N}(0,1), t} [\| \epsilon - \epsilon_{\theta}(z_t, t, \tau_{\theta}(y)) \|_2^2] \quad (2.6)$$

This objective comes directly from Rombach *et al.* (2022) [3] and is analogous to that used in DDPMs. Intuitively, it trains ϵ_{θ} to “denoise” z_t by recovering the injected Gaussian noise. When the model is optimal, one can iteratively apply it (subtracting the predicted noise) to transform a pure-noise latent $z_T \sim \mathcal{N}(0, I)$ into a

coherent latent z_0 , which is then decoded to an image.

In the above objective, the variables have the following roles:

- x is the original (clean) image, and $z_0 = E(x)$ is its latent encoding.
- y is the conditioning input (e.g. a text description), with $\tau_\theta(y)$ its neural embedding.
- t is a discrete timestep index (sampled uniformly from $\{1, \dots, T\}$, indicating how much noise has been added.
- $\epsilon \sim N(0, I)$ is the actual Gaussian noise used at step t .
- z_t is the noisy latent after t steps of forward diffusion: $z_t = \sqrt{\bar{\alpha}_t} \cdot z_0 + \sqrt{1 - \bar{\alpha}_t} \cdot \epsilon$
- $\epsilon_\theta(z_t, t, \tau_\theta(y))$ is the neural network's prediction of the noise given the noisy latent, the time t , and the condition.
- The loss $\| \epsilon - \epsilon_\theta(\cdot) \|^2$ thus measures how well the model predicts the true noise [3]

By minimizing this loss over the data distribution x , y and noise ϵ , the model learns to reverse the forward diffusion: starting from random noise, repeatedly subtracting the predicted ϵ from z_t will recover $z_{t-1} \approx z_0$. Finally, applying the decoder D to z_0 yields the generated image.

(2) Architecture Overview

Stable Diffusion assembles three main components-autoencoder, denoising UNet, and text conditioning via CLIP and cross-attention-into an end-to-end image synthesis pipeline [3], [20]. Each plays a distinct role:

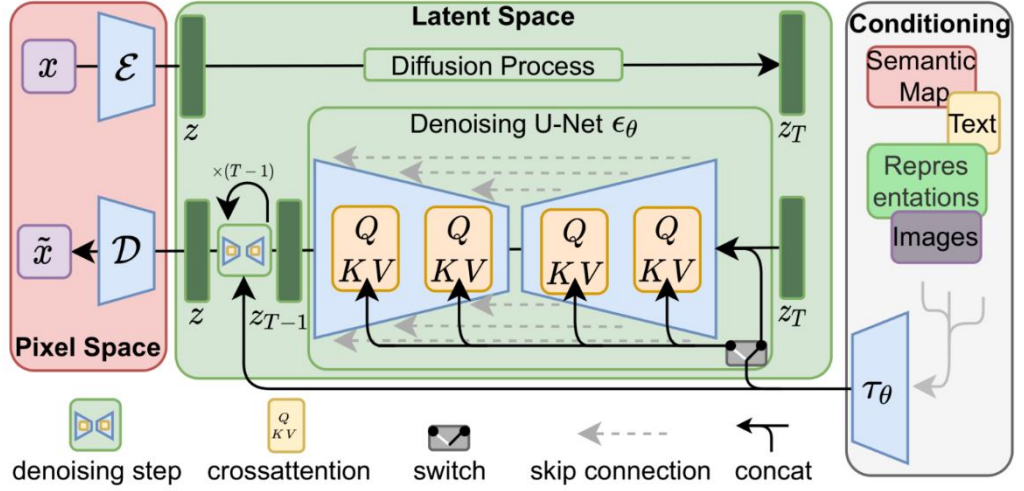


Figure 2. Architecture of Latent Diffusion Model with Cross-Attention

Conditioning (High-Resolution Image Synthesis with Latent Diffusion Models [2])

The image synthesis pipeline begins with a variational autoencoder that compresses input images from high-resolution pixel space into a compact latent space. Given an image $x \in \mathbb{R}^{H \times W \times 3}$, the encoder maps it to a lower-dimensional latent representation $z = \varepsilon(x) \in \mathbb{R}^{h \times w \times c}$, where typically $h, w \ll H, W$. This latent encoding retains the semantic structure of the image while discarding redundant pixel-level details. Operating the diffusion process in this compressed latent space reduces computational and memory requirements significantly. During both training and inference, the forward noising process is applied to the encoded latent z_0 , and the final denoised latent $\hat{z}_0$ is decoded back into the image domain via a single pass through the decoder D , yielding $\hat{x} = D(\hat{z}_0)$. This architectural design makes the diffusion model more tractable and scalable without compromising visual fidelity [3].

The core of the denoising process is handled by a U-Net that operates on the noisy latent variables z_t . This U-Net incorporates a time embedding to condition its predictions on the specific timestep t , which determines the noise level present in the input. The time embedding is represented through sinusoidal or learnable positional encodings and injected into the network alongside the latent features. At each denoising step, the U-Net predicts the noise component ϵ that was added during the forward process. By progressively subtracting these estimates, the model reconstructs the clean latent representation. The network’s encoder-decoder structure with skip connections allows it to model both local and global image patterns, supporting

effective denoising across multiple spatial scales [3].

To ensure semantic alignment between generated images and user-specified prompts, the model incorporates text conditioning through a CLIP-based embedding pipeline [20]. A textual prompt y is encoded into a sequence of embeddings $\tau_\theta(y)$, which are injected into the denoising U-Net using cross-attention layers at specific stages of the network.

In the cross-attention mechanism, the latent image features act as queries Q , while the keys K and values V are derived from the text embeddings. The attention operation:

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{Q \cdot K^T}{\sqrt{d}} \right) \cdot V \quad (2.7)$$

Allows the image features to selectively focus on relevant aspects of the prompt, effectively guiding the generation toward semantically consistent outputs. This dynamic interaction between visual and textual information ensures that the model can produce images that faithfully reflect the intent and specificity of the input prompt [20].

(3) Sampling Process Visualization

While the previous section explained the architecture of Stable Diffusion, this part illustrates how the model actually generates an image at inference time. Starting from random noise $z_t \sim N(0, I)$, the model iteratively denoises it through the trained UNet to obtain a clean latent representation, which is then decoded into the final image $x = D(z_0)$ [3].

To generate an image, the model starts from Gaussian noise $z_t \sim N(0, I)$ and iteratively denoises through the UNet using:

$$z_{t-1} = \frac{1}{\sqrt{\alpha_t}} \cdot \left(z_t - \frac{\beta_t}{\sqrt{1-\alpha_t}} \cdot \epsilon_\theta(z_t, t) \right) + \sigma_t \cdot z, \quad z \sim N(0, I) \quad (2.8)$$

Where:

z_t : noisy latent at step t

$\alpha_t = 1 - \beta_t$: retention factor at step t

$\bar{\alpha}_t = \prod_{i=1}^t \alpha_i$: cumulative product

σ_t : standard deviation of added noise

This is repeated for T steps until z_0 , which is decoded to image space: $\hat{x} = D(z_0)$

(4) Motivation and Role in my Project

In my project, I utilize a pre-trained Stable Diffusion model as a starting point to generate personalized images. Rather than training the model from scratch, I fine-tune only the concept embedding, allowing the model to associate a new personalized token (e.g., the name of a person) with a specific visual identity [4], [5]. This approach has several advantages:

- **Efficiency:** Leveraging a pre-trained model avoids extensive training while maintaining high image quality.
- **Modularity:** The personalized embedding can be trained separately and plugged into the original model.
- **Scalability:** It allows the generation of personalized results using minimal reference data.

In conclusion, the Stable Diffusion model forms the core of my generative framework, offering a balance between performance and flexibility, which enables me to focus on personalizing the model for specific identity-aware generation.

2.1.2 Text Embeddings

In text-to-image diffusion models, text embeddings play a foundational role as the conditioning signal that steers the image synthesis process toward semantically aligned outputs. Given a user-defined natural language prompt y , the sentence is first tokenized into a sequence of tokens $\{t_1, t_2, \dots, t_n\}$, where each token represents a word or subword unit depending on the tokenizer employed—typically Byte Pair Encoding (BPE) or WordPiece tokenization [52]. This tokenization allows the model to flexibly handle vocabulary variations, out-of-vocabulary terms, and morphological richness.

(1) Embedding Lookup and CLIP Encoding

Each token t_i is mapped to a learnable embedding vector $v_i \in R^d$, where $d=1024$ in the case of Stable Diffusion v2.1 [3]. These raw token embeddings are then processed by a pretrained CLIP text encoder ε , which converts them into contextualized vector representations that capture not only individual word semantics but also sentence-level meaning. The resulting output is:

$$c(y) = [\varepsilon(v_1), \varepsilon(v_2), \dots, \varepsilon(v_n)] \in R^{n \times 1024} \quad (2.9)$$

Here, $\varepsilon(v_i)$ denotes the contextualized embedding for token v_i , and the sequence $c(y)$ constitutes the textual conditioning input to the diffusion model. These embeddings are kept frozen during inference and standard image generation tasks, preserving the knowledge learned during large-scale pretraining on image-text pairs [6].

(2) Role in Diffusion Process

The text embeddings $c(y)$ are injected into the diffusion model-typically via cross-attention layers inside a U-Net architecture-at multiple spatial resolutions. At each denoising timestep t , the latent representation z_t is conditioned on both the time embedding and the textual embedding $c(y)$, which guides the reverse diffusion trajectory toward a sample that reflects the content of the prompt y [3], [37].

This conditioning mechanism enables the model to align the visual structure of the output with the semantic content of the input prompt. The text encoder, trained jointly with CLIP [6], is particularly adept at capturing compositional semantics, e.g., “a cat wearing sunglasses” or “a sunset in a cyberpunk city,” enabling highly coherent and stylistically faithful generation.

(3) Properties of Text Embeddings

The strength of CLIP-based text embeddings lies in their ability to:

- Encode both local and global context within a prompt.
- Map semantically similar prompts (e.g., “a cute puppy” and “a small dog”) to nearby regions in embedding space.
- Enable zero-shot generalization to novel prompts, styles, or object

combinations.

In practice, this embedding space serves as a shared latent space with image embeddings, facilitating strong vision-language alignment in multimodal tasks.

(4) Toward Personalized Embeddings

While CLIP embeddings are general-purpose and effective for common prompts, they are not sufficient when attempting to personalize the generation process to unseen subjects-such as a specific person, object, or artwork. In such cases, the textual embedding must be augmented or replaced with a learned personalized embedding.

These personalized embeddings are optimized to represent the identity and attributes of a target subject, using only a small number of reference images. Critically, they retain compatibility with the standard conditioning pathway through the CLIP-based encoder and cross-attention mechanism.

In summary, text embeddings provide the linguistic foundation for conditioning text-to-image diffusion models. The integration of pretrained CLIP text encoders ensures that prompts are interpreted with semantic richness and contextual nuance. This mechanism not only drives general-purpose generation but also supports the integration of learned personalized embeddings, allowing for identity-aware and contextually guided image synthesis.

2.1.3 Textual Inversion method

Modern text-to-image generation models such as Stable Diffusion are capable of synthesizing rich and diverse images from arbitrary text prompts. However, when the goal is to personalize the generation process-by embedding a specific object, individual, or visual style-the challenge becomes more complex. Retraining the entire model for each new concept is computationally expensive and prone to catastrophic forgetting, where the model loses general knowledge acquired during pretraining.

Textual Inversion, introduced by Gal et al. (2022), addresses this problem with a simple yet powerful approach [4]. Rather than modifying the model's architecture or parameters, it injects a new concept into the text embedding space of a frozen model by:

1. Introducing a new pseudo-word token, denoted as $\langle S^* \rangle$, into the tokenizer's vocabulary.
2. Optimizing a single embedding vector for $\langle S^* \rangle$ so that, when used in prompts like “a photo of $\langle S \rangle$ ” * , the model generates images resembling a small set of 3–5 reference images.
3. Freezing the entire model, except for the $\langle S^* \rangle$ embedding, which is updated using gradient descent on a reconstruction loss.

Once trained, $\langle S^* \rangle$ can be seamlessly inserted into any textual prompt-e.g., “an oil painting of $\langle S \rangle$ ”, “ $\langle S \rangle$ on the beach”-allowing users to synthesize the concept in novel styles and settings with high fidelity and flexibility [4].

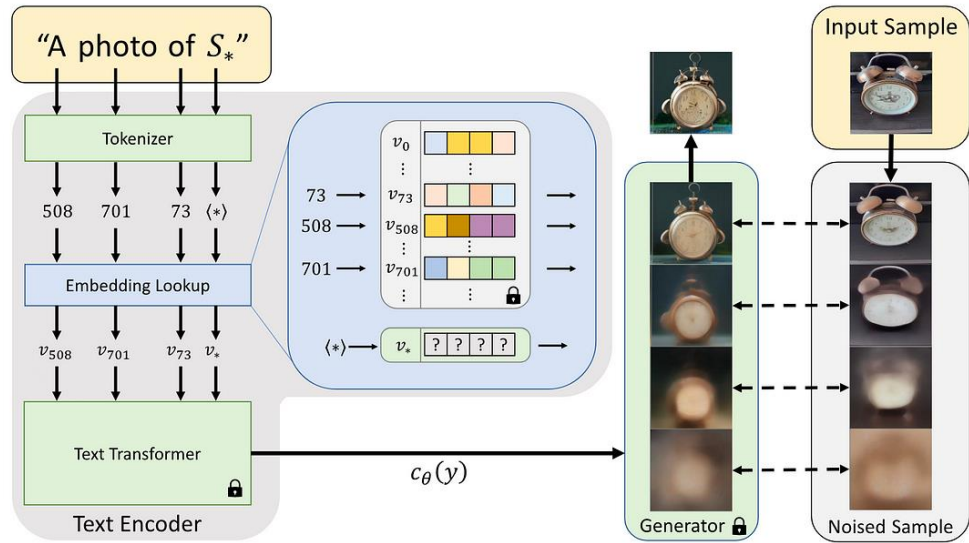


Figure 3. Outline of the text-embedding and inversion process (Textual Inversion [4])

2.1.3.1 Technical Details of Textual Inversion

The process begins by defining a new token $\langle S^* \rangle$ that is not present in the model's vocabulary. This token is initialized with an embedding vector-either randomly or by copying from an existing generic word such as “object” or “sculpture.” The tokenizer and embedding matrix are extended to include this new token [4].

The goal is to optimize the embedding vector v^* for $\langle S^* \rangle$ such that the model generates images matching the reference concept. At each training step [4]:

- A reference image x is encoded into a latent representation $z_0 = E(x)$

- Gaussian noise is added to simulate the forward diffusion process:

$$z_t = \sqrt{\alpha_t} \cdot z_0 + \sqrt{1 - \alpha_t} \cdot \epsilon, \epsilon \sim N(0, I) \quad (2.10)$$

- A text prompt containing $\langle S^* \rangle$ (e.g., “*A photo of $\langle S \rangle$* ”*) is randomly selected from a template pool.
- The model predicts the noise $\hat{\epsilon}$ using the prompt embedding $c_\theta(\text{prompt})$, which contains v^* .
- The reconstruction loss is computed as:

$$L_{rec} = E_{x,t,\epsilon} [\| \epsilon - \epsilon_0(z_t, t, c_\theta(\text{prompt})) \|^2] \quad (2.11)$$

- Only the embedding vector v^* is updated using gradients; all other model weights remain frozen.

After sufficient optimization (typically $\sim 5,000$ steps), the token $\langle S^* \rangle$ encodes the visual identity of the target concept. It can now be used in various prompts to generate personalized images in new contexts, such as:

- “ $\langle S \rangle$ made of lego”*
- “*A da Vinci sketch of $\langle S \rangle$* ”*
- “ $\langle S \rangle$ in a carpet stamp”*

Thanks to the model’s rich linguistic and visual priors, it can reason over prompts containing $\langle S^* \rangle$ and compose them into diverse scenes while preserving the concept’s identity [4].

2.1.3.2 Strengths and Extensions

- Strengths:
 - No fine-tuning of the main model: Only one embedding vector is trained.
 - Data-efficient: Works well with as few as 3–5 reference images.
 - Highly composable: The learned token can be used across varied prompts and contexts.
- Proposed Extensions: To improve effectiveness and stability, several

extensions have been proposed [4]:

- Multi-token embeddings: Using 2–3 tokens instead of one to better capture fine-grained details.
- Regularization: Penalizing deviation from a generic descriptor to prevent semantic drift.
- Per-image tokens: Assigning additional tokens per image to disentangle shared and unique attributes across the dataset.

Overall, Textual Inversion is a foundational method for personalizing large-scale text-to-image models without retraining or modifying their architecture. By learning a concept-specific token directly in the text embedding space, it enables efficient, modular, and semantically meaningful generation. Its simplicity and effectiveness make it a key building block for downstream applications in creative design, virtual identity synthesis, personalized content generation, and beyond [4].

2.1.4 The Cross Initialization method

One of the most critical challenges in personalized text-to-image generation is how to effectively encode the unique visual identity of a subject into the language space of a pre-trained diffusion model such as Stable Diffusion. A common solution is Textual Inversion, in which I introduce a new, learnable token (e.g., [V]) and optimize its embedding such that the model generates images closely resembling a target individual when prompted with that token [4].

However, traditional Textual Inversion suffers from two major limitations:

- Ineffective initialization: New token embeddings are often initialized randomly, leading to poor convergence as the model struggles to capture the subject’s unique identity [5].
- Semantic drift: As training progresses, the learned embedding may deviate significantly from the original semantic space of language, making it challenging to combine the personalized concept with broader text prompts. This misalignment also complicates prompt composability and editing [5].

To overcome these limitations, I adopt a novel approach called Cross

Initialization, proposed by Pang et al. (2023). This method provides a semantically meaningful and structurally aligned initialization for the new concept token, improving both identity fidelity and editability, which are essential to my project's goals [5].

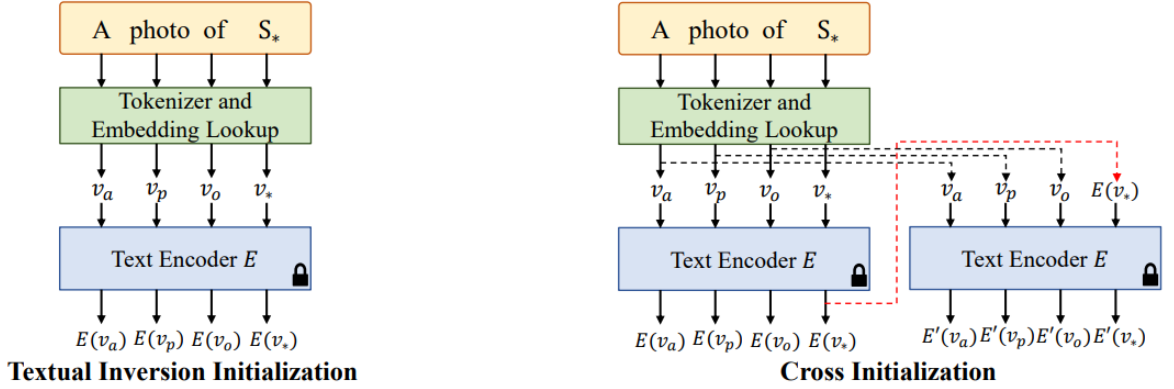


Figure 4. Comparison of Textual Inversion Initialization and Cross Initialization techniques. (Cross Initialization for Personalized Text-to-Image Generation paper [5])

2.1.4.1 Core Mechanism of Cross Initialization

In my project, one of the most important objectives is to generate accurate, identity-preserving images of a person given only minimal supervision—typically a single reference image and their name. The concept embedding must not only capture their appearance but also support flexible combinations with different prompts (e.g., "a portrait of [V] in the forest" or "a painting of [V] smiling") [5]. Therefore, I adopt Cross Initialization as a foundational method for concept embedding optimization.

The Cross Initialization method involves two main stages: Text-Based Initialization and Regularized Optimization. These stages are designed to ensure that the personalized embedding is both semantically accurate and stable, leading to better image generation results.

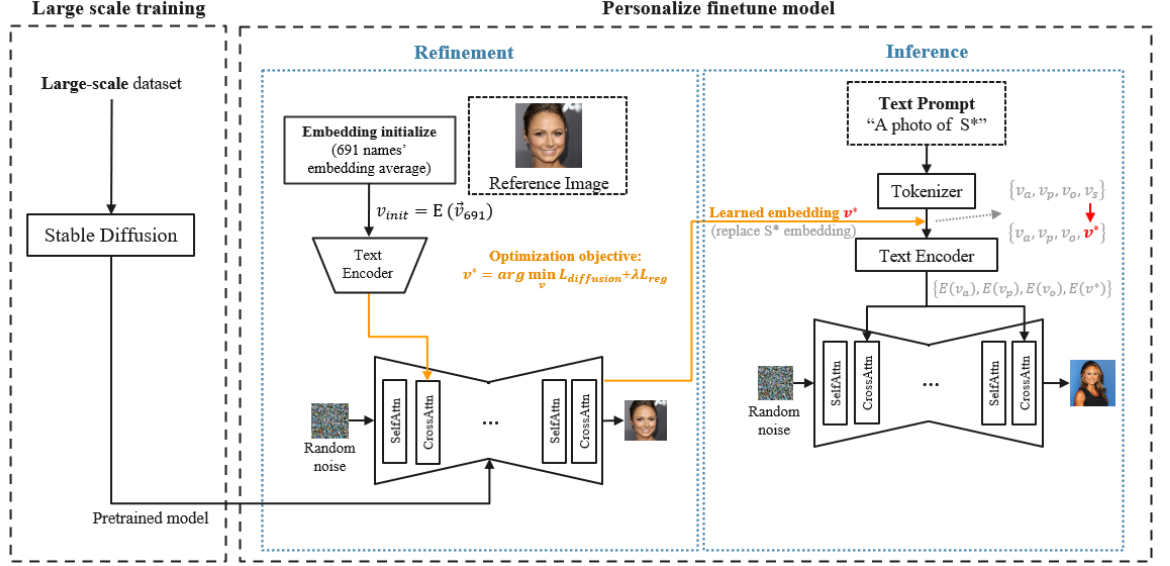


Figure 5. Cross Initialization Method structure

(1) Text-Based Initialization:

(a) Constructing the Celebrity Name List:

A list of 691 celebrity names is collected to ensure broad diversity across different domains, such as actors, musicians, and athletes. Each name is represented by two tokens: the first token corresponds to the first name (e.g., “Emma”), and the second token corresponds to the last name (e.g., “Watson”). This dataset is curated to capture a semantically rich and diverse representation of human identity in the language embedding space [5].

(b) Text Encoding with CLIP:

Each celebrity name is then processed through the CLIP text encoder, which converts the name into a high-dimensional vector. CLIP (Contrastive Language–Image Pretraining) is a powerful model trained on large-scale image-text pairs, capable of generating semantically meaningful textual representations. The output of the CLIP encoder is a semantic embedding for each name, capturing both linguistic and conceptual information associated with human identity [6], [53].

(c) Mean Embedding Computation:

Once all celebrity names are encoded, the mean embedding $\bar{v}_{691}$ is computed from the vectors of the 691 names. This embedding represents a generic human identity in the language space. The mean embedding captures the commonalities among the various celebrity names, including shared linguistic and semantic features

related to human identity [5].

A total of 691 wellknown names are used to form an embedding set $C = \{v_1, \dots, v_m\}$. The mean textual embedding is calculated as :

$$\bar{v}_{691} = \frac{1}{m} * \sum_{i=1}^m v_i \quad (2.12)$$

where $m = 691$ and v_i denotes the embedding of the i -th name. Since each name consists of two tokens (first and last names), two separate averages are computed: $\bar{v}_{691}^f$ and $\bar{v}_{691}^l$. These are concatenated to form the final mean embedding [5]:

$$\bar{v}_{691} = [\bar{v}_{691}^f, \bar{v}_{691}^l] \quad (2.13)$$

(d) Generating the New Token Embedding via Cross Initialization:

Based on the analysis in Section 4.1 [5], the initialization of the personalized token embedding should meet two criteria: (1) it should be close to the final learned embedding, and (2) it should provide a rough semantic approximation of the target concept.

Two key observations motivate the proposed Cross Initialization strategy:

- After optimization, the learned embedding becomes similar to the output of the text encoder.
- When the text encoder's output is used as its own input, the diffusion model can generate an image nearly identical to the original, without requiring encoder fine-tuning.

These effects are attributed to the LayerNorm and self-attention layers in the text encoder, which normalize and rotate embeddings such that they converge toward a common direction and scale. Therefore, Cross Initialization sets the initial embedding as the encoder output of the mean vector:

$$v_{init} = E(\bar{v}_{691}) \quad (2.14)$$

This initialization ensures the new embedding starts from a semantically stable and meaningful position in the language space. As shown in the experiments in [5], the learned embedding remains close to this initialized state throughout training, thereby mitigating semantic drift, accelerating convergence, and improving both identity preservation and prompt consistency during image generation.

(2) Regularized Optimization

To further refine the embedding and prevent it from drifting too far from the initialized value, a regularization term is added to the loss function. This term helps ensure that the learned embedding retains its connection to the original semantic space.

After initialization, the personalized embedding is optimized using two loss components:

- **Reconstruction Loss $L_{\text{diffusion}}$:** This is the standard loss used in latent diffusion models (LDMs), which measures how accurately the model can reconstruct the noise that was added during the forward diffusion process. It drives the model to learn a denoising function that faithfully reflects the target visual concept in generated outputs.
- **Regularization Loss L_{reg} :** Encourages the learned embedding to stay close to the mean embedding $\bar{v}$, which is constructed by averaging the embeddings of 691 well-known names. The region around this central embedding is assumed to represent the semantic subspace corresponding to the concept of individuals. High editability is expected when the learned embedding lies close to this subspace. The regularization is implemented by minimizing the L_2 distance between the learned embedding v and the central embedding $\bar{v}$ [5]:

$$L_{\text{reg}} = ||v - v_{\text{init}}||_2^2 \quad (2.15)$$

- The overall optimization objective becomes:

$$v^* = \arg \min_v. L_{\text{diffusion}} + \lambda. L_{\text{reg}} \quad (2.16)$$

where:

- $L_{\text{diffusion}}$ promotes accurate noise prediction and reconstruction,
- L_{reg} enforces semantic proximity to a human-identity prior, and
- λ is a scalar coefficient that balances the influence of regularization in the overall loss function [5].

The regularized loss function ensures that the embedding does not deviate significantly from the semantic space defined by the mean embedding. This improves the stability of the learning process and allows the model to converge faster, as the optimization is more constrained by the semantic alignment.

2.1.4.2 Impact and Outcomes

The Cross Initialization method significantly enhances the stability and semantic fidelity of the concept embedding initialization. By using the mean embedding of a large set of celebrity names and regularizing the optimization process, this approach mitigates the issues of ineffective initialization and semantic drift commonly seen in traditional Textual Inversion. The result is a more reliable and faster learning process, with the personalized token retaining its alignment with the broader language space.

This initialization method lays a strong foundation for the successful personalization of text-to-image generation, enabling the model to generate high-quality, identity-preserving images based on minimal supervision [5].

2.2 Proposed method

2.2.1 Analysis

In the following section, we analyze key limitations of name-based embedding initialization-specifically the `celeb_names_cross_init` approach, which is representative of current cross-initialization methods. Highlighting these weaknesses helps clarify the requirements for a more effective and personalized embedding strategy.

2.2.1.1 Motivation for the Proposed Method

In traditional cross-initialization methods-such as `celeb_names_cross_init`-the personalized token embedding is initialized by averaging the embeddings of a set of reference texts, typically the names of 691 well-known celebrities, which are pre-

encoded using CLIP. This approach provides a semantically rich starting point, as celebrity names often carry high-level information such as gender, ethnicity, style, or profession. However, despite being simple and effective during the early personalization stage, this method suffers from significant limitations when the goal is to model a specific identity in detail:

- **Lack of true personalization:** Name-based embeddings cannot encode fine-grained visual features such as facial structure, skin tone, hairstyle, or expression, all of which are crucial for faithfully reconstructing identity.
- **Limited representational space:** Since the embedding is derived from a fixed set of 691 celebrities, when the input reference image belongs to a novel individual outside this set, the averaged embedding fails to capture semantic cues specific to that person. This is especially problematic in applications demanding highly accurate personalization.
- **Low accuracy in recreating specific appearances:** When the task is to generate an image of a particular person from a single reference image, the name-based initialization is insufficient to capture the complexity of individual appearance, often leading to semantic drift in the generated output.

These limitations underscore the need for a more expressive and flexible embedding initialization strategy—one that can effectively bridge the semantic intent from language with the visual specificity of the reference image. This motivates the proposed method, `fusion_cross_init`, which incorporates both text and image information to construct a personalized token embedding that is better aligned with the user’s intent and the individual’s true appearance.

2.2.1.2 Theoretical Foundation

(1) Back ground: Cross Initialization via Text Embeddings

In `celeb_names_cross_init`, a list of names (e.g., ["Emma Watson", "Brad Pitt"]) is used as a weak proxy for the identity to be learned. The process includes:

- Tokenizing each name using a CLIP-compatible tokenizer [6].
- For each token position (e.g., first token, second token...), average the embeddings across all names.
- Remove special tokens ([BOS], [EOS]) and pad to reach a fixed length

(e.g., 75).

- Format the sequence as [BOS] + v_{init} + [EOS] + [PAD] $\rightarrow$ input of length 77 [6].
- Pass through the CLIP Text Encoder to obtain the final refined v_{init} .

This approach is text-only and relies on the assumption that the selected names are semantically and visually aligned with the target concept-which may not always hold.

(2) Theory of Fusion Cross Init

Recent advancements in multimodal learning-such as CLIP [6], BLIP [56], and Flamingo [54] have demonstrated that integrating visual and textual signals significantly enhances semantic alignment between input modalities and model outputs. These systems leverage the complementary nature of language and vision to form rich, context-aware embeddings that are both semantically expressive and visually grounded.

The underlying rationale is based on the strengths of each modality:

- Text encoders (e.g., from CLIP or BLIP) provide high-level semantic intent, capturing abstract concepts, descriptive attributes, and contextual cues from the input text. For example, phrases like “*a smiling woman in traditional dress*” encode both identity attributes and scene-level semantics.
- Vision encoders extract low-level appearance features, such as color, texture, facial structure, and lighting conditions, which are crucial for identity preservation and fine-grained visual details.

By fusing these modalities at the embedding level, I can construct representations that benefit from the abstract understanding of language while maintaining fidelity to visual characteristics. Such joint embeddings are crucial in personalized generation tasks, where both who (identity) and what (semantic content) must be preserved.

To effectively combine these two types of features, I adopt an advanced fusion strategy based on a Transformer architecture. Specifically, I design a Fusion Transformer module, in which the text embeddings (derived from celebrity names) and the visual features (extracted from a reference image) are provided as two

separate input streams. This module employs a cross-attention mechanism to learn the correlations between semantic information and visual characteristics, thereby enabling the model to dynamically learn how to optimally fuse the two modalities. The output is a rich, expressive initial embedding, which is then used as the personalized token to guide the image generation process.

Using a Transformer-based approach offers several advantages:

- It allows the model to capture complex, non-linear relationships between linguistic and visual information, rather than relying on simple linear mixing.
- The flexible attention mechanism enhances semantic-visual alignment, leading to more coherent outputs.
- It improves the model's generalization capability across diverse textual prompts and reference images.

In this project, this fusion principle forms the core of the improved embedding initialization method. By integrating name-based semantic embeddings with reference image features through the Fusion Transformer, the proposed method achieves better alignment between user intent and generated outputs, producing images that are both identity-faithful and semantically accurate.

2.2.2 Enhanced Token Initialization via Text-Image Fusion (Fusion Cross Init)

The `fusion_cross_init` method is designed to improve the initialization of the personalized token embedding by combining both textual semantics and visual features from reference inputs. A stronger initialization enables the model to converge faster and better preserve identity during fine-tuning, particularly when text prompts alone lack descriptive richness.

This method builds upon the traditional `celeb_names_cross_init`, which initializes the token embedding by averaging the CLIP text embeddings of 691 celebrity names. However, such embeddings are generic and may not represent the unique appearance of a novel individual.

To address this, I propose an extended approach where visual features from a reference image are extracted and fused with the semantic text-based embedding, as shown in the “Customized Refinement” stage of the diagram:

1. **Textual Prior:** First, a semantic prior $V_{init}=E(\vec{v}_{691})$ is computed by averaging the embeddings of 691 celebrity names using a CLIP text encoder.
2. **Visual Features:** Simultaneously, features are extracted from the reference image using a visual encoder.
3. **Fusion Transformer Module:** These two modalities-semantic and visual-are then combined using a Fusion Transformer module, marked as step (2) in the diagram. This module employs cross-attention to learn complex relationships between the two feature types and produces a multi-modal enriched token embedding.
4. **CLIP-Compatible Token Formatting :** To align with CLIP’s text encoder, the fused text-image embeddings are reformatted into a 77-token sequence, where each embedding serves as a surrogate for a new personalized token; the full sequence is then passed to CLIP as input embeddings to produce a refined vector v_{init} , which is used to initialize the personalized concept.

The resulting embedding v_{init} serves as the personalized token that is used both during the fine-tuning stage and inference (by replacing the placeholder token s^*). This method ensures the initialized embedding is semantically grounded and visually aligned, thus facilitating more accurate and identity-preserving image generation from textual prompts.

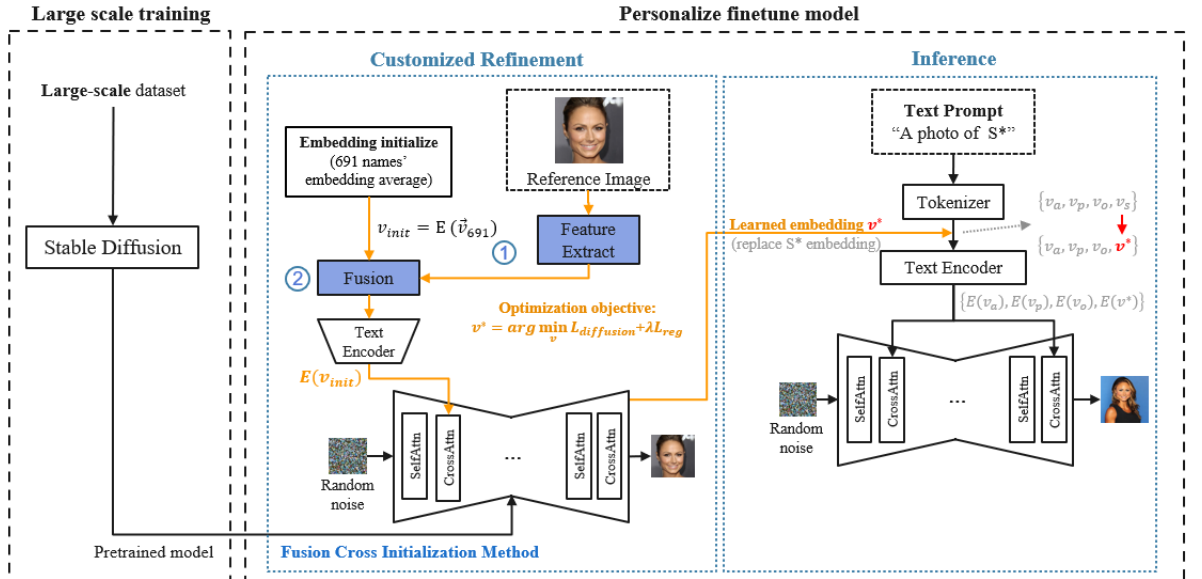


Figure 6. Fusion Cross Initialization Method structure

2.2.2.1 Input Data:

The `fusion_cross_init` method requires several specific inputs. First, `celeb_path` is the path to a text file containing a list of celebrity names, which provides initial semantic information. `ref_image_path` is the path to a reference image, serving as the source of the individual's visual features. The `fusion_weight` parameter is a float between 0 and 1 that determines the weighting between textual and visual information during the fusion process. Additionally, the method uses CLIP-compatible modules such as the tokenizer and `text_encoder` to encode the inputs. Finally, `n_column` specifies the number of tokens used during embedding initialization (default is 75), ensuring compatibility with CLIP's 77-token input format.

2.2.2.2 Detailed Structure:

(1) Step 1: Text Processing (Cross initialization)

Based on the Cross Initialization method, the personalized embedding initialization process begins by tokenizing a list of celebrity names using CLIP's tokenizer. This step decomposes each name into smaller subword units (tokens), enabling the CLIP model to handle names with different lengths and structures more effectively. For instance, the name "Adam Sandler" is tokenized into [BOS, adam, sand, ler, EOS], where [BOS] and [EOS] are special tokens indicating the beginning and end of the sequence.

Next, the special tokens [BOS] and [EOS] are explicitly removed to prevent positional bias in the resulting embeddings, as these tokens do not carry any semantic information specific to the identity. This ensures that only meaningful content-bearing tokens are retained for further processing.

Then, the remaining tokens are grouped by their respective positions within each name. Specifically, the first tokens after [BOS] across all names form group v_0 , the second tokens form group v_1 , and so on-until a predefined maximum number of positions is reached. This positional grouping is essential, as it allows the model to learn shared semantic features from corresponding parts of different names. For example, v_0 may include first names like "emma", "brad", or "tom", capturing consistent identity-level characteristics.

Each token group v_i is then encoded into vector embeddings using the CLIP text encoder, a pre-trained component designed to map natural language tokens into a high-dimensional semantic space. As a result, each token is represented as a 1024-dimensional vector, encoding contextual and linguistic meaning.

Finally, the embeddings in each positional group are averaged to produce a single representative vector for that position. This averaging operation enhances common semantic traits shared across different identities and smooths out individual variances. The final result is a text embedding matrix of shape $[n_{column}, 1024]$, where n_{column} refers to the number of token positions retained (typically 75, aligning with CLIP's 77-token input format, excluding the [BOS] and [EOS] tokens), and 1024 is the embedding dimensionality output by the CLIP text encoder.

** The dimension 1024 is used because the CLIP Text Encoder outputs 1024 dimensional embeddings [6].*

(2) Step 2: Image Processing (Visual Encoding)

The image processing pipeline in the Cross Initialization method consists of three main stages: image preprocessing, feature extraction, and projection to text embedding space—each playing a specific role in aligning visual and textual representations for personalized embedding generation.

The process begins with standardizing the reference image to make it compatible with CLIP's vision encoder: The input image is first converted to RGB format with a fixed resolution of 224×224 pixels, ensuring consistent dimensions and color channels expected by the CLIP model. The image is then normalized using CLIP's predefined mean and standard deviation values. This normalization step brings pixel values into a standardized range that matches the CLIP model's training distribution, improving inference accuracy. After normalization, the image is converted into a tensor and reshaped to include a batch dimension, enabling it to be processed correctly by the neural network.

Once the image tensor is ready, it is passed through the CLIP Vision Encoder, which has been trained to extract rich, localized visual features from the input image. The encoder produces two types of embeddings: a sequence of patch-level embeddings of shape $[num_patches, 768]$, where each embedding corresponds to a

spatial region of the image and captures local visual semantics; and a pooled embedding of shape [768], typically obtained through mean pooling or by using the special [CLS] token, which encodes a global representation of the entire image. Instead of relying solely on patch-level features, both patch and pooled features are included in the final representation. The pooled embedding provides complementary global context, which enhances identity preservation and semantic alignment when fusing with text features later in the pipeline. [6].

Since CLIP's text embeddings are 1024-dimensional while the image encoder outputs 768-dimensional features, a dimensionality mismatch arises. To address this, a linear projection layer is applied to map image embeddings from 768 to 1024 dimensions, ensuring compatibility in the fused multimodal representation. This step allows image features to be semantically aligned with text features before fusion.

(3) Step 3: Fusion Transformer Module-A Deep Integration Mechanism Between Text and Image

To enable effective fusion between image and text features, the Fusion Cross Init method utilizes a Transformer-based architecture instead of relying solely on a linear projection to align dimensions. The goal of the FusionTransformer architecture is to generate a text embedding that is deeply integrated with information from corresponding images. This is particularly important in tasks such as Textual Inversion, Personalized Text-to-Image Generation, or Multimodal Alignment, where text tokens need to accurately reflect the visual characteristics of the associated images.

(a) Concatenate sequences:

The text embeddings of shape $[N, \text{dim}]$ and image embeddings of shape $[M, \text{dim}]$ are concatenated along the sequence length dimension (i.e., the first dimension in PyTorch when `batch_first=True`), resulting in a Combined Sequence of shape $[N+M, \text{dim}]$. More specifically:

- Inputs:
 - Text Embeddings $T \in R^{N.d}$: Consists of N tokens, including the words/phrases from the prompt and, most importantly, the placeholder token to be trained. Each token is encoded into a d -dimensional vector

using a model such as the CLIP Text Encoder.

- Image Embeddings $I \in R^{M.d}$: Consists of M vectors, which are the output of the CLIP Vision Encoder, projected from the original 768-dimensional space to the 1024-dimensional text space for alignment. These vectors can represent either image patches or a global image representation.
- Concatenation of Embeddings : Once the text and image embeddings have been encoded and projected into the same semantic space of dimension d , the next step is to integrate them into a single unified representation. This is done by concatenating the two sequences along the sequence length dimension (dimension 0 in PyTorch when `batch_first=True`):

$$C = \text{Concat}(T, I) \in R^{(N+M).d} \quad (2.17)$$

Where:

- $T \in R^{N.d}$ is the input text embedding sequence of N tokens (including placeholder tokens).
- $I \in R^{M.d}$ is the image embedding sequence, which may represent image patches or pooled visual features, already projected to the same dimensionality d as text.
- $C \in R^{(N+M).d}$ is the combined sequence of embeddings.

By concatenating them into a single sequence, I allow the subsequent Transformer encoder layers to model cross-modal dependencies directly. This setup enables every token (from either modality) to attend to every other token, fostering bidirectional interactions and deeper semantic fusion between text and image information. This strategy is especially effective in tasks like Personalized Text-to-Image Generation or Textual Inversion, where the placeholder token's meaning should be shaped both by textual context and the corresponding image characteristics.

This Combined Sequence is then passed through a Transformer Encoder, which consists of multiple layers (`num_layers`) of standard Transformer encoder blocks.

(b) Transformer Encoder:

The purpose of using a Transformer Encoder is to allow the model to learn complex cross-modal interactions between linguistic (text) and visual (image) information from the combined sequence $C \in R^{(N+M).d}$. This enables the text embeddings to not only reflect language semantics, but also to deeply incorporate visual cues from the corresponding image. Specifically, Transformer layers help the model:

- Enable each token (whether text or image) to attend to all other tokens in the combined sequence, learning inter-modal relationships.
- Produce text embeddings that have been enriched with image information, which is essential for tasks such as Textual Inversion or Multimodal Alignment.

Each Transformer Encoder layer consists of two main components that sequentially refine the joint representation of text and image modalities.

The first component is the Multi-Head Self-Attention (MHSA) mechanism. The primary objective of MHSA is to enable each token-whether originating from text or image-to attend to every other token in the input sequence, thereby learning both contextualized and cross-modal relationships. This attention mechanism dynamically weights contextual information, allowing tokens to adaptively integrate relevant semantic or visual signals depending on the context.

The MHSA mechanism starts with a concatenated sequence of token embeddings from both text and image:

$$C \in R^{(N+M).d}$$

where N is the number of text tokens, M is the number of image tokens, and d is the embedding dimension.

The input embedding sequence C is linearly transformed into three distinct spaces: Query (Q), Key (K), and Value (V). This is done by multiplying C by learnable weight matrices W^Q , W^K , and W^V , respectively:

$$Q = C \cdot W^Q, K = C \cdot W^K, V = C \cdot W^V \quad (2.18)$$

Each of these projections plays a distinct and crucial role in the attention mechanism. Queries (Q) can be thought of as the “questions” posed by each token, seeking to gather relevant information from other tokens in the sequence. Keys (K) serve as “addresses” or identifiers that specify the content and characteristics of each token's information. Values (V) contain the actual data or information that will be aggregated to form new token representations.

This separation into queries, keys, and values allows the model to dynamically compute how much attention or weight each token should give to every other token, by measuring the similarity between the query of one token and the key of another. The result is a context-sensitive weighting system that lets tokens flexibly incorporate semantic or visual cues from the entire input sequence.

The projection matrices $W^Q, W^K, W^V \in R^{d \times d_k}$ are trainable parameters that transform the input embedding dimension d into a lower-dimensional subspace d_k , which is often set as $d_k = \frac{d}{h}$, where h is the number of attention heads. This division enables parallel attention heads to specialize and focus on different aspects or types of features, improving the model's ability to learn complex patterns.

Attention scores are calculated through a scaled dot product between queries and keys. The raw dot products are divided by the square root of the key dimension $\sqrt{d_k}$ to mitigate issues caused by large magnitude values, which can slow down or destabilize training. The scaled scores are then passed through a softmax function to generate a probability distribution that assigns attention weights to each token:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{Q \cdot K^T}{\sqrt{d_k}}\right) \cdot V \quad (2.19)$$

This produces a weighted sum of value vectors, where weights reflect the relevance of each token to the token issuing the query. Thus, each token's representation becomes a dynamic aggregation of information from the entire input sequence, enabling rich, context-dependent feature learning that adapts to the semantics or visual features present.

Multi-Head Self-Attention (MHSA) further enhances this process by

computing several attention operations in parallel. Each of the h heads independently applies attention with its own learned projections Q_i, K_i, V_i capturing diverse subspace representations:

$$head_i = \text{Attention}(Q_i, K_i, V_i) \quad (2.20)$$

The outputs of all heads are concatenated and linearly transformed by a matrix W^O to combine the multi-faceted attention information into a single unified representation:

$$\text{MHSA}(C) = \text{Concat}(head_1, \dots, head_h) \cdot W^O \quad (2.21)$$

This multi-head design allows the model to attend to information at multiple levels of abstraction simultaneously—for example, capturing local fine-grained interactions, global dependencies, cross-modal relationships, or different semantic aspects. It improves the expressiveness and robustness of the learned embeddings, leading to better performance in tasks requiring complex reasoning or multimodal fusion.

Moreover, by integrating information from both text and image tokens in a joint sequence C , this attention mechanism enables deep cross-modal contextualization. Each token can flexibly attend to relevant tokens regardless of their modality, fostering a more aligned and semantically rich joint representation crucial for downstream tasks such as personalized text-to-image generation.

After computing the Multi-Head Self-Attention (MHSA), the output is combined with the original input through a Residual Connection. Residual connections are a crucial technique in deep neural networks designed to mitigate the vanishing gradient problem and facilitate more effective training. By allowing the original input signal to bypass one or more layers unchanged, the model focuses on learning the residual—i.e., the difference between the input and desired output—leading to faster convergence and reduced risk of overfitting.

Mathematically, the residual connection is expressed as:

$$C' = C + \text{MHSA}(C) \quad (2.22)$$

Here, C denotes the input embedding sequence, and $\text{MHSA}(C)$ represents the computed attention output. This additive operation preserves the fundamental characteristics of the original embeddings while incorporating the newly learned contextual information derived from attention.

Subsequently, the combined output passes through Layer Normalization (LN). Layer Normalization is a normalization technique applied independently to each sample, contrasting with Batch Normalization which operates across batches. LN stabilizes the input distribution to subsequent layers, enhancing gradient stability during training and improving overall model convergence and performance.

The LN operation can be mathematically described as:

$$\text{LN}(C') = \frac{C' - \mu}{\sigma} \odot \gamma + \beta \quad (2.23)$$

where μ and σ are the mean and standard deviation of the components within C' , and γ , β are learnable parameters that rescale and shift the normalized values.

The synergy of Residual Connection and Layer Normalization offers multiple advantages: it preserves original information, prevents degradation through deep layers, and enables learning of more complex and meaningful representations. Especially in multimodal tasks such as text-image fusion, these mechanisms effectively maintain and enhance interactions between different modalities, crucial for generating rich embeddings that comprehensively integrate contextual and visual information.

In summary, MHSA combined with Residual Connection and Layer Normalization forms a robust framework for learning deep, multi-dimensional, and cross-modal representations. This enables the Transformer not only to comprehend individual modalities but also to intricately capture the interplay between them.

(c) Feed-Forward Network (FFN):

After self-attention integrates contextual information across the combined sequence C , the updated representation C' is passed through a position-wise Feed-

Forward Network (FFN). This component introduces non-linearity and further refines each token's representation independently, without considering interactions with other tokens.

Mathematically, the FFN applies the following transformation to each token embedding $x \in R^d$:

$$\text{FFN}(x) = \text{GELU}(x \cdot W_1 + b_1) \cdot W_2 + b_2 \quad (2.24)$$

Here:

- $W_1 \in R^{d_{ff} \cdot d}$ and $W_2 \in R^{d \cdot d_{ff}}$ are learnable weight matrices.
- $b_1 \in R^{d_{ff}}$ and $b_2 \in R^d$ are bias vectors.
- d_{ff} is the hidden dimension of the intermediate layer in FFN, typically set to 4 times the model dimension d . For example, when $d=1024$, then $d_{ff}=4096$.
- GELU (Gaussian Error Linear Unit) is used as the activation function, chosen for its smoothness and empirical effectiveness in large-scale language and vision models.

Unlike the attention mechanism which operates across token positions, the FFN is applied identically and independently to each token. This makes it a position-wise transformation, allowing the model to increase its capacity for learning non-linear mappings and abstract features at each location in the sequence.

(d) Residual Connection and Normalization

Following the FFN, a residual connection is applied to encourage gradient flow and preserve information from earlier layers:

$$x_{res} = x + \text{FFN}(x) \quad (2.25)$$

Then, a Layer Normalization is applied:

$$x_{out} = \text{LayerNorm}(x_{res}) \quad (2.26)$$

This design pattern (Residual $\rightarrow$ Non-linearity $\rightarrow$ Residual $\rightarrow$ Normalization)

is crucial for stabilizing training and improving convergence in deep networks. It ensures that the representations are well-behaved statistically, reducing internal covariate shift.

Summary of One Transformer Encoder Layer: A complete Transformer Encoder layer processes the combined sequence $C \in R^{(N+M).d}$, where N is the number of text tokens and M is the number of image tokens, in the following stages:

1. Self-Attention Mechanism: Computes contextual relationships across all tokens, capturing interactions between textual and visual features.
2. Cross-Modal Fusion: Allows the model to dynamically integrate semantic content across modalities (e.g., a word attending to an image patch).
3. Feed-Forward Network: Applies a position-wise non-linear transformation to refine each token's embedding.
4. Residual Connections and Layer Normalization: Ensures stable learning and mitigates degradation in very deep models.

After repeating this structure across L encoder layers, the final output representation is:

$$C(L) = \text{EncoderOutput} \in R^{(N+M).d} \quad (2.27)$$

(4) Step 4: CLIP-Compatible Token Formatting

The main goal of this step is to construct a token sequence with a fixed length of exactly 77 tokens - the standard input length expected by CLIP's Text Encoder. Ensuring the token sequence has this fixed length is crucial for maintaining compatibility and effectively leveraging CLIP's semantic representation capabilities.

- The constructed sequence takes the form:

$$[\text{BOS}] + \text{fused_emb}_{0 \rightarrow n-1} + [\text{EOS}] + [\text{PAD} \times (77 - n - 2)] \quad (2.28)$$

Where:

- [BOS] (Beginning of Sequence): A token marking the start of the sequence, helping the model identify where the input begins.

- fused_emb_i: The fused embeddings sequence, where each element acts as a surrogate token representing a newly personalized concept, e.g., pseudo tokens like <new_custom>_v₀, <new_custom>_v₁, etc.
- [EOS] (End of Sequence): A token marking the end of the sequence, signaling the model where the input ends.
- [PAD]: Padding tokens used to fill the sequence to the fixed length of 77 tokens. The number of padding tokens is 77−n−2, accounting for 1 BOS, 1 EOS, and n personalized tokens.
- Importance of This Formatting: The fixed length of 77 tokens is crucial because CLIP was trained on inputs of this length, ensuring compatibility and optimal performance. Special tokens like BOS and EOS define the contextual boundaries of the input, enhancing the model's understanding. A key innovation lies in the use of personalized fused_emb tokens, which replace native vocabulary tokens to represent new, unseen concepts—enabling the model to learn and encode personalized information more effectively. Padding is also applied to maintain uniform sequence lengths, supporting efficient batch processing and consistent architecture.

Embedding Refinement Process: The constructed token sequence is passed into the text_encoder as input_embeds=..., meaning instead of providing token IDs, the model receives the pre-fused embeddings directly as input. The output is a refined embedding $v_{init} \in R^{n \cdot 1024}$, where n is the number of personalized tokens (e.g., 2 tokens). This v_{init} serves as the initialized embedding for the new personalized concept, providing a specialized, robust representation suitable for subsequent training or personalized image generation tasks.

2.2.2.3 Illustration of Workflow

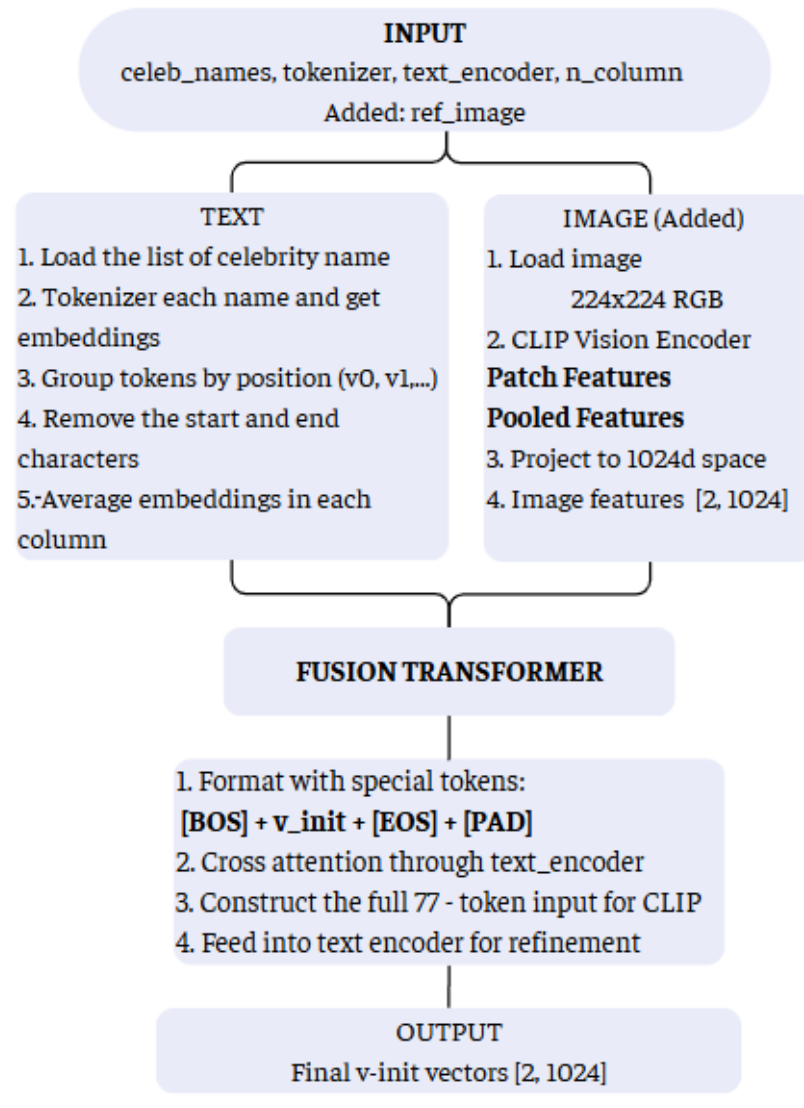


Figure 7. Flowchart of the Fusion Cross Initialization operational process

CHAPTER 3

EXPERIMENTAL RESULTS & EVALUATION

This chapter presents the experimental framework used to assess the practical effectiveness of the proposed approach. It outlines the evaluation setup, datasets, and metrics employed to analyze performance in various personalization scenarios.

Rather than relying solely on traditional name-based initialization, the proposed method integrates both visual and textual cues during the embedding construction phase. To validate its advantages, a series of experiments are conducted under controlled conditions, with a baseline comparison against existing techniques.

The evaluation process covers multiple dimensions of model behavior, with a particular focus on how well the system can generate identity-consistent and semantically accurate outputs from limited reference data. Results are presented using both objective quantitative metrics and subjective visual inspection to provide a comprehensive assessment of the method's strengths and limitations. The findings in this chapter serve as evidence for the proposed method's capacity to support more effective and generalizable personalization in text-to-image generation tasks.

3.1 Experiment

The primary objective of this experiment is to evaluate the effectiveness of the proposed Fusion Cross Initialization (Fusion-CI) method in enhancing identity preservation and semantic alignment in personalized text-to-image generation. The evaluation focuses on three key aspects: identity consistency, prompt editability, and optimization efficiency.

This study is guided by the following research questions:

1. Does the proposed Fusion-CI method more effectively preserve individual identity in generated images, particularly under diverse or challenging prompts?
2. Does Fusion-CI improve semantic alignment between the generated image and the input prompt compared to the Cross Initialization (CI) baseline?

To address these questions, I conduct a comparative analysis under controlled experimental settings, evaluating against the state-of-the-art personalization method: Cross Initialization (CI). Both quantitative metrics and qualitative visual assessments are employed.

Through these evaluations, I aim to demonstrate that incorporating both textual and visual information during the embedding initialization phase leads to more accurate, identity-preserving, and semantically coherent personalized image synthesis.

3.1.1 Implementation Details and Training Configuration

In my experiments, I utilize Stable Diffusion v2.1 [45], a widely available open-source model, as the base for the personalized text-to-image generation task. I choose this model due to its strong performance in generating high-quality images and its compatibility with existing personalization techniques. All images are generated at a fixed resolution of 512×512 pixels, balancing image quality and computational cost.

I initialize the optimization process with a regularization hyper-parameter λ set to 10^{-5} , which helps maintain a balance between identity preservation and alignment with text descriptions during the embedding optimization.

My personalization pipeline follows a one-shot learning setting, where only a single reference image of the target identity is provided. The entire experiment is conducted on a NVIDIA RTX 2080 GPU, with a batch size of 4, a learning rate of 0.000625, and gradient accumulation over 3 steps. This setup allows me to simulate a larger effective batch size and improve training stability without exceeding GPU memory limits.

The optimization is performed over 500 steps, during which the personalized embedding is iteratively refined. This number of steps is empirically chosen to ensure sufficient alignment between the identity and text prompts while minimizing overfitting or identity drift. After training, the final personalized embedding is used to guide the diffusion model for text-conditioned image generation.

To ensure reproducibility, I fix all random seeds. The training process is implemented in PyTorch with mixed-precision training to optimize memory efficiency. Additional implementation details such as prompt selection and token initialization strategy are provided in the appendix.

3.1.2 Evaluation Setup

For the evaluation, I use test images from the CelebA-HQ dataset [57, 58]. The text prompts are primarily adapted from [17] and [47], with minor modifications for better semantic alignment. I compare my proposed method with a baseline personalization technique: Cross Initialization [5]. All methods are implemented under the one-shot personalization setting, where a single reference image is provided for each identity.

For the quantitative evaluation, I apply each method to a large number of images, including both randomly selected individuals and samples from the CelebA-HQ test set. Two evaluation metrics are used:

- **Identity Similarity:** This metric measures the consistency between the generated image and the reference identity. It is computed using ArcFace [59], a pretrained face recognition model that extracts facial features and quantifies identity preservation based on cosine similarity.
- **Prompt Similarity:** This metric evaluates how well the generated image aligns with the textual description. I calculate the CLIP score between the generated image and its corresponding text prompt to assess semantic alignment.

Together, these metrics offer a robust and interpretable framework for evaluating both the fidelity and consistency of personalized text-to-image synthesis results. I exclude prompts with stylization content (e.g., "in Van Gogh style", "in Pixar animation") from the identity similarity evaluation, since ArcFace is trained on real-world facial images and may not accurately assess identity under heavy stylistic alterations.

3.1.2.1 Prompt Similarity Metric – CLIP Score

As text-to-image generation models become increasingly sophisticated, evaluating their performance requires more than just assessing visual quality. A central challenge is measuring how well the generated image semantically aligns with the input text prompt—a capability not captured by traditional metrics.

Classic evaluation metrics such as:

- Inception Score (IS): Assesses clarity and diversity of generated images.
- Fréchet Inception Distance (FID): Measures the statistical distance between generated and real image distributions.

Focus primarily on visual realism and diversity but fail to determine whether the content of the generated image actually matches the prompt. For instance, if the prompt is “*a red apple on a table*” and the model outputs a realistic image of an orange, IS and FID might still score it well—despite being semantically incorrect.

CLIP Score, proposed by Hessel et al. (2021) [60], was designed to address this gap. It provides a semantic-aware metric that evaluates whether the generated image reflects the intended meaning of the prompt, by leveraging the CLIP model trained on large-scale image-text pairs [6].

(1) Technical Foundation

CLIP (Contrastive Language-Image Pretraining) learns to embed both images and texts into a shared semantic space, such that matched image-text pairs are close while mismatched ones are far apart. CLIP Score uses this capability to quantify how well a generated image semantically aligns with its corresponding text prompt.

Given a text prompt T and a generated image I , the score is computed as follows:

1. The prompt is encoded into a feature vector: $f_T = \text{CLIP}_{\text{text}}(T)$
2. The image is encoded into a feature vector: $f_I = \text{CLIP}_{\text{image}}(I)$
3. Normalize both vectors to unit length
4. Cosine similarity is computed between the normalized vectors:

$$\text{CLIPScore}(T, I) = \frac{f_T \cdot f_I}{\|f_T\| \cdot \|f_I\|} \quad (3.1)$$

- Where:
 - f_T : Feature vector of the text prompt, encoded by CLIP's text encoder.
 - f_I : Feature vector of the generated image, encoded by CLIP's image encoder.
 - $\cdot$: Dot product (cosine numerator).
 - $\|f_T\|$ and $\|f_I\|$: Norms (magnitudes) of the respective vectors.

The resulting score typically ranges from 0 to 1, where a higher score indicates stronger semantic alignment between the text and the image.

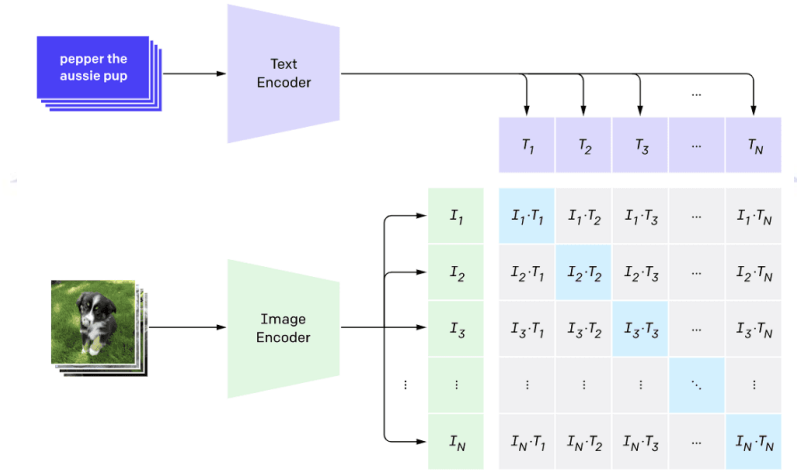


Figure 8. Example of an evaluation matrix model: CLIP-I, CLIP-T

(Source: [OpenAI](#))

(2) Practical Applications

CLIP Score is widely adopted in both academic and industrial settings for:

- Benchmarking text-to-image models: It is used to compare different versions of models like DALL·E, Imagen, or Stable Diffusion.
- Auxiliary loss in training: Some approaches integrate CLIP Score into the training loss function to explicitly optimize for semantic alignment [20].
- Image reranking: When multiple images are generated from the same prompt, CLIP Score can be used to select the most relevant one.

Its flexibility and generalization make it especially valuable for large-scale automated evaluations.

(3) Complementary Metrics

To obtain a comprehensive assessment, CLIP Score is often used in conjunction with other metrics, such as:

- FID: for realism and distribution similarity.
- IS: for diversity.
- Human evaluation: for nuanced qualitative judgment.
- Perceptual Path Length (PPL): for measuring smoothness of interpolation.

(4) Conclusion

CLIP Score represents a significant advancement in the evaluation of text-to-image generation systems. By measuring semantic compatibility between image and text using the pretrained CLIP model, it offers a robust, automated, and scalable method to assess alignment quality-an essential factor in modern multimodal AI. However, to obtain a comprehensive understanding of model performance, CLIP Score should be combined with other metrics such as FID and human evaluation.

3.1.2.2 Identity Similarity – ArcFace-Based Evaluation

(1) Evaluation Objective

In the context of personalized image generation from a single reference image (one-shot personalization), a key evaluation criterion is the model's ability to preserve the visual identity of the subject in the generated outputs. The generative model must not only produce visually appealing images but also maintain consistent identity features of the person in the reference image.

To assess this, ArcFace [\[59\]](#) - a pretrained face recognition model-is employed to compare a single reference image against a set of generated images, measuring how well the identity is retained across the outputs.

(2) Evaluation Methodology

Face Embedding Extraction:

- The reference image is passed through ArcFace to extract the identity vector f_{ref}

- Each generated image is also passed through ArcFace to obtain its corresponding identity vector f_{gen_i}

Identity Similarity Computation:

- The cosine similarity is computed between the reference vector and each generated vector:

$$\text{Sim}_i = \frac{f_{\text{gen}_i} \cdot f_{\text{ref}}}{\|f_{\text{gen}_i}\| \cdot \|f_{\text{ref}}\|} \quad (3.2)$$

- Where:
 - f_{gen_i} : Feature vector of the i -th generated image.
 - f_{ref} : Feature vector of the reference image.
 - $\cdot$: Dot product between the two vectors.
 - $\|f\|$: Norm (length) of the vector f , its magnitude
- The final Identity Similarity score is the average similarity over all generated images:

$$\text{Identity Score} = \frac{1}{N} \sum_{i=1}^N \text{Sim}_i \quad (3.3)$$

- Where:
 - N : Total number of generated images.
 - Sim_i : Cosine similarity between the i -th generated image and the reference image (as calculated in the previous formula).
 - The score reflects the overall identity preservation quality of the generation method.

Interpretation: The similarity score ranges from 0 to 1, where higher values indicate better identity preservation.

(3) Usefulness and Significance

Using ArcFace in this setup provides the following advantages:

- Objective evaluation: ArcFace focuses solely on facial identity, independent of image aesthetics.
- Effective for one-shot settings: It enables robust comparison between a single input and multiple outputs.
- Highly suitable for personalized generation tasks, such as avatar creation, face restoration, or identity-conditioned generation.

(4) Limitations

As ArcFace is trained on real human faces: Generated images that are heavily stylized (e.g., cartoons, sketches, artistic paintings) should be excluded from identity similarity evaluation to ensure reliability of the metric.

ArcFace-based identity similarity measurement is a robust and accurate tool for evaluating personalized face generation. It provides a reliable comparison between a single reference image and a batch of generated outputs, ensuring that personalization maintains the subject's core identity. This makes it a crucial metric alongside others like CLIP Score in evaluating multimodal generative models.

P F I E V

3.2 Results

- Input:



Refinement steps: 500

Train batch size: 4

Learning rate: 0.000625

Inference steps: 100

Figure 9. Input image: Person 1

- Output:

Table 2. Experimental Results for Person 1: Cross Initialization Method & Fusion

Cross Initialization Method

Prompt	Cross Init	Fusion
“A photo of S* person”		
“A photo of S* is eating birthday cake”		
“A photo of S* is playing the guitar”		
“A photo of S* with angry expression”		

- Input:



Refinement steps: 500
Train batch size: 4
Learning rate: 0.000625
Inference steps: 100

Figure 10. Input image: Person 2

- Output:

Table 3. Experimental Results for Person 2 : Cross Initialization Method & Fusion
Cross Initialization Method

Prompt	Cross Init	Fusion
“A photo of S* person”		
“A photo of S* is talking with Barack Obama”		
“A photo of S* is playing the guitar”		
“A photo of S* is coding”		

3.2.1 Qualitative Evaluation

The visual results presented in the two tables illustrate a direct comparison between the Cross Initialization (CI) and Fusion Cross Initialization methods. The images were generated based on various types of prompts, including basic portrait descriptions, specific actions (e.g., eating cake, playing guitar), facial expressions (e.g., angry), social interactions (e.g., talking with Barack Obama), and professional contexts (e.g., coding).

Across both tested subjects, the Fusion Cross Initialization method demonstrates significantly better identity preservation, especially in prompts that

require both accurate semantic interpretation and consistent facial features. While Cross Initialization occasionally produces blurry images or unstable identity representations, Fusion Cross Initialization maintains sharper and more consistent facial identity across a range of visual contexts.

In terms of semantic alignment with the prompts, Fusion Cross Initialization also exhibits higher flexibility. In scenarios such as “S* is playing the guitar” or “S* is coding,” the generated images not only preserve the subject’s identity but also correctly reflect the described action or setting. In contrast, Cross Initialization tends to produce repetitive or less accurate representations of the intended content, particularly in activity-based or interaction-heavy prompts.

Notably, in cases involving facial expression editing (e.g., “S* with angry expression”), Fusion Cross Initialization is the only method that successfully reflects emotional changes without distorting facial identity. Meanwhile, Cross Initialization often results in expressions that are either inconsistent or difficult to interpret.

Overall, the qualitative results show that Fusion Cross Initialization not only ensures accurate identity preservation but also excels in composing scenes, actions, and emotions more effectively than the traditional Cross Initialization method. Its stability and high image quality across diverse scenarios confirm its strong potential for real-world personalized text-to-image generation under rich and complex input conditions.

3.2.2 Quantitative evaluation

The quantitative evaluation used two metrics: CLIP Score to assess semantic alignment between the prompt and the generated image, and ArcFace Similarity to measure identity preservation relative to the reference image.

Table 4 presents the evaluation results of two personalized text-to-image generation methods, using ArcFace and CLIP Score as evaluation metrics, based on the input images and prompts provided in Table.

For each prompt, 50 images are generated and evaluated. The CLIP score is computed between each generated image and its corresponding prompt, while ArcFace cosine similarity is computed between each generated image and the single

reference image. The scores are averaged per prompt to ensure statistical stability and fair comparison. The results are presented in the following table.

Higher values indicate better performance. In each pairwise comparison between CI and Fusion, the better-performing method can be identified directly by comparing the numerical values.

Table 4. Comparison of Evaluation Metrics Between Two Methods: Cross Initialization Method & Fusion Cross Initialization Method

Image	Prompt	Method	Identity similarity	Prompt similarity
Person 1	“A photo of S* person”	CI	0.4200	0.2583
		Fusion	0.4315	0.2602
	“A photo of S* is eating birthday cake”	CI	0.3201	0.2962
		Fusion	0.3512	0.3063
	“A photo of S* is playing the guitar”	CI	0.2653	0.2729
		Fusion	0.2801	0.2635
Person 2	“A photo of S* person”	CI	0.3343	0.2825
		Fusion	0.3401	0.2842
	“A photo of S* is talking with Barack Obama”	CI	0.3800	0.2584
		Fusion	0.3901	0.2625
	“A photo of S* is playing the guitar”	CI	0.0333	0.2828
		Fusion	0.0354	0.2925
	“A photo of S* is coding”	CI	0.2522	0.2773
		Fusion	0.3015	0.2770
		CI	0.2743	0.2573
		Fusion	0.2950	0.2766

The table 4 presents a detailed comparison of identity preservation (via ArcFace similarity) and prompt-image alignment (via CLIP score) between the Cross Initialization (CI) and Fusion Cross Initialization (Fusion) methods across two subjects and various text prompts.

Several key observations can be made:

1. **Overall Improvement in Identity Similarity:** In most prompts, the Fusion method achieves higher identity similarity scores compared to CI. For example, for “*A photo of S is eating birthday cake*”* (Person 1), Fusion yields a score of 0.3512, higher than CI’s 0.3201. This trend suggests that Fusion Cross Init better preserves personal identity, likely due to its more effective integration of visual and textual features.
2. **Better Prompt Alignment in Most Cases:** The CLIP score, which evaluates semantic alignment with the prompt, is generally improved or maintained with the Fusion method. Notably, for “*S is eating birthday cake*”* and “*S is talking with Barack Obama*”*, the prompt similarity with Fusion is consistently higher than CI. This demonstrates Fusion’s advantage in maintaining the meaning of the input prompt in the generated image.
3. **Stable Performance Across Diverse Prompts:** The Fusion method performs robustly across various types of prompts, from simple descriptions (“*S* person*”) to more complex scenes (“*S* is coding*” or “*S* with angry expression*”). This consistency highlights its generalization ability, especially important for personalized generation tasks involving unseen or challenging text descriptions.
4. **Few Exceptions:** In a few cases, Fusion’s improvement is marginal or slightly underperforms in prompt similarity (e.g., “*S is playing the guitar*”* for Person 1). This may be due to model sensitivity to specific visual contexts or limitations in text-image fusion under certain scenarios.

Table 5. Extended Comparison of Evaluation Metrics Between Cross Initialization and Fusion Cross Initialization Methods

Method	Identity similarity	Prompt similarity
Cross Initialization	0.3015	0.2742
Fusion Cross Initialization	0.3325	0.2789

Next, to further evaluate the effectiveness of the proposed method, I used 30 reference images, each combined with five different types of prompts to ensure semantic diversity. For every image–prompt pair, I generated 50 output images for each method (CI and Fusion-CI), resulting in a total of 15,000 generated images ($30 \times 5 \times 2 \times 50$).

For identity similarity (ArcFace), I calculated the cosine similarity between the ArcFace embedding of each generated image and its corresponding reference image. I then averaged the similarity scores over 50 generated samples for each method and prompt, and further averaged the results across all 30 reference images to obtain the final evaluation score.

For prompt similarity (CLIP Score), I computed the CLIP Score between each generated image and its corresponding prompt. The same two-level averaging was applied: first over 50 images per method/prompt, and then over all 30 reference images.

This two-stage averaging ensures statistical stability and fairness in comparing the two methods. Table 5 presents the final evaluation results, showing that the proposed Fusion Cross Initialization method consistently outperforms the original Cross Initialization, particularly in its ability to preserve personal identity features.

Overall, these findings indicate that the Fusion Cross Initialization method clearly outperforms the traditional Cross Initialization approach in both identity preservation and semantic alignment with the input prompt. The integration of CLIP-aware mechanisms during initialization significantly enhances visual consistency and

semantic fidelity in the generated images, thereby validating the effectiveness of this method for personalized text-to-image generation tasks.

3.2.3. Runtime & Computational Efficiency

Although the proposed Fusion Cross Initialization (Fusion-CI) method significantly improves identity preservation and semantic alignment, evaluating its feasibility for real-world deployment requires further analysis of computational efficiency and runtime performance. Below is a preliminary assessment:

(1) Training Time and Memory Usage

The Fusion-CI method enhances the baseline personalization process by introducing an additional Fusion Transformer module, which is designed to integrate visual and textual features in a more effective and semantically aligned manner. This module enables deeper interaction between the semantic information from the textual prompt and the visual characteristics extracted from reference images, thus supporting more accurate identity preservation. While this architectural enhancement introduces slight additional complexity to the model, the overall computational overhead during training remains minimal and acceptable for practical usage.

- In terms of training time, Fusion-CI demonstrates excellent efficiency. Experimental comparisons under the same conditions (i.e., identical dataset, model settings, and hardware environment) show that the training time increases by only about 2.5% compared to the standard Cross Initialization method. For instance, performing 500 refinement steps using Cross Initialization takes approximately 20 minutes, whereas the same number of steps with Fusion-CI requires only 20 minutes and 30 seconds. This marginal increase reflects the efficiency of the added fusion mechanism and confirms its practicality for research and application environments with limited computational resources.
- Regarding memory consumption, the Fusion-CI method requires approximately 5–10% more GPU memory during training. This increase is primarily attributed to the multi-head attention operations in the Fusion

Transformer, which simultaneously processes both the text embeddings (e.g., from a celebrity name) and the visual embeddings (e.g., from reference images). The memory impact is most noticeable during the fusion stage, where these two modalities are aligned through a sequence of attention-based interactions. Despite this, the model remains lightweight enough to be trained on mid-tier GPUs such as the NVIDIA RTX 2080 without the need for additional hardware optimization.

Overall, the experimental results show that Fusion-CI achieves its intended functional improvements with minimal trade-offs in training efficiency, making it a viable solution for identity-preserving personalization in text-to-image generation tasks.

(2) Hardware and Inference Speed

All experiments in this study were conducted using a NVIDIA RTX 2080 GPU with 8GB of VRAM, a commonly available consumer-grade GPU. This configuration was selected to reflect a realistic hardware environment that balances computational capability and accessibility for practical deployment scenarios.

- The average image generation time using the proposed Fusion-CI method was approximately 6.2 seconds per image at a 512×512 resolution. This performance is within an acceptable range for personalized content generation tasks, especially in offline or near-real-time applications.
- Notably, the inference phase remains highly efficient, since the personalization is injected through a learned token embedding rather than requiring any modification of the core parameters of the Stable Diffusion model. As a result, the method avoids the overhead associated with fine-tuning large model weights and maintains compatibility with existing inference pipelines. This token-based personalization strategy ensures that the computational cost during inference stays comparable to the baseline generation process, even when personalized embeddings are used.

- Overall, the system demonstrates a good balance between identity preservation accuracy and inference efficiency, suggesting that Fusion-CI can be integrated into practical applications without significantly increasing computational demand.

(3) Optimization Suggestions for Deployment

To make the Fusion-CI method more accessible for broader deployment-including in near real-time systems or on mid-tier hardware-some lightweight optimization strategies may be considered in future work, even though the current implementation already shows minimal computational overhead.

- Pruning (optional): While the current model runs efficiently, further reducing redundant attention heads or layers in the Fusion Transformer could help scale the system to lower-end hardware, if necessary.
- Quantization (for future deployment): Converting the Fusion module to lower-precision operations (e.g., FP16 or INT8) may further reduce memory usage and improve inference speed, particularly on devices with limited computational resources.

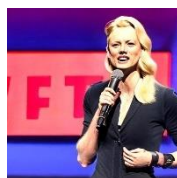
Although these optimizations are not strictly required in the current setting-as Fusion-CI already demonstrates competitive training and inference performance-they offer promising directions for making the method more scalable and adaptable in large-scale or edge-based deployments.

3.3 Further Experiments with Different Inputs

To further validate the robustness and generalization ability of the proposed method, I conducted additional experiments using inputs that differ from the original evaluation setup. These variations include different subjects, alternative textual prompts, and diverse visual contexts. The aim is to assess how well the model adapts to new or challenging conditions while maintaining identity preservation and semantic alignment.



Real sample



"S* giving a TED talk on stage"



"S* reading a book in the library"



"S* sitting and watching sunset"



"S* painting a landscape on canvas"



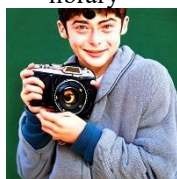
"S* dancing in the rain"



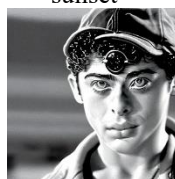
Real sample



"S* watching fireworks by a lake"



"S* holding a vintage camera"



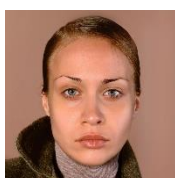
"S* in a noir film scene"



"S* walking a dog in autumn leaves"



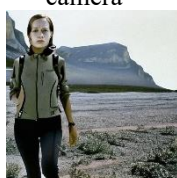
"S* wrapped in a cozy blanket, drinking tea"



Real sample



"S* standing next to a rocket"



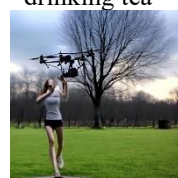
"S* hiking alone in the mountains"



"S* teaching a class of students"



"S* fixing a bike on the street"



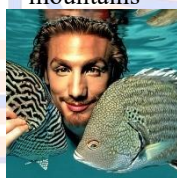
"S* flying a drone in the park"



Real sample



"S* riding a horse through a snowy field"



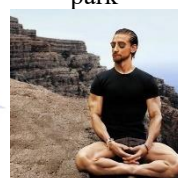
"S* underwater with a school of fish"



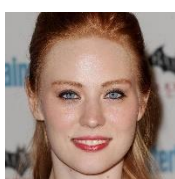
"S* as a clay sculpture"



"S* running a marathon with a number tag"



"S* meditating on top of a cliff"



Real sample



"S* as a medieval knight in armor"



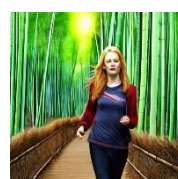
"S* singing on a big concert stage"



"S* with a cool face"



"S* doing a backflip on the beach"



"S* walking through a bamboo forest"



Real sample



"S* drawing graffiti on a wall"



"S* planting a tree with children"



"S* covered in paint after art class"



"S* as an astronaut on Mars"



"S* cooking noodles at a food stall"



Figure 11. Extended Evaluation Results on Additional Test Cases of the Fusion Cross Initialization Method

To further assess the generalization capability of the Fusion Cross Initialization method, we conducted extended experiments on 10 additional individuals, using a total of 50 diverse text prompts. These prompts cover a wide range of categories, including expressions, actions, backgrounds, interactions, and artistic styles, thereby offering a more comprehensive evaluation of the method's adaptability.

The results, summarized in Table 4, demonstrate that the Fusion Cross Initialization method consistently maintains strong performance across different subjects and prompt types. It achieves high identity similarity scores while preserving

semantic alignment with the prompts, even in more challenging or previously unseen scenarios.

This confirms the method’s robustness, versatility, and scalability, indicating its suitability for real-world personalized text-to-image generation tasks involving greater diversity in user inputs. Overall, the extended evaluation shows very promising and stable results, further validating the effectiveness of the proposed approach.



CONCLUSIONS

Personalized text-to-image generation is a highly promising topic in the field of artificial intelligence and computer vision, with broad applications in content creation, entertainment, fashion, medical imaging, and identity reconstruction. However, achieving both strong identity preservation and accurate semantic alignment with the input prompt remains a difficult challenge. It requires the model to effectively balance visual fidelity and linguistic understanding. In this project, I tackled this challenge by improving the embedding initialization stage—an essential part of the personalization pipeline for diffusion models like Stable Diffusion.

Specifically, I proposed a new method called Fusion Cross Initialization (Fusion-CI), which combines information from two distinct modalities: visual features extracted from a reference image, and semantic features encoded by the CLIP text encoder. Unlike existing methods such as Textual Inversion or traditional Cross Initialization, which rely solely on text features, Fusion-CI uses a soft fusion mechanism to generate a more balanced and expressive initialization embedding that captures both identity and semantic intent.

The method was implemented on Stable Diffusion v2.1 and trained on an NVIDIA RTX 2080 GPU using optimized settings (batch size, learning rate, gradient accumulation, etc.). For quantitative evaluation, two main metrics were used: identity similarity, measured by ArcFace embeddings, and prompt similarity, assessed using CLIP Score. The experimental results showed that Fusion-CI consistently achieved higher identity similarity compared to Cross Initialization, particularly for prompts involving facial expressions—a challenging aspect of identity preservation. While prompt similarity scores (CLIP Score) between the two methods were relatively close, Fusion-CI still maintained competitive performance, with slight improvements in some prompt types.

These findings suggest that Fusion-CI offers a meaningful improvement in terms of maintaining personal identity, which is essential in applications where consistent appearance across varied prompts is crucial. By introducing a more

expressive initialization embedding, the model avoids the common pitfalls of prior methods, such as overfitting to the reference image or ignoring key elements of the prompt. Moreover, the proposed method demonstrates strong computational efficiency. Despite the architectural enhancement introduced by the Fusion Transformer, training time only increases marginally (approximately 2.5%), and memory usage remains within a manageable range for mid-tier GPUs such as the RTX 2080. The inference speed also remains comparable to the baseline, as personalization is applied via a learned token embedding without modifying the core model weights. These factors suggest that Fusion-CI is not only effective but also practical for real-world deployment, even in resource-constrained environments.

These findings suggest that Fusion-CI offers a meaningful improvement in terms of maintaining personal identity, which is essential in applications where consistent appearance across varied prompts is crucial. By introducing a more expressive initialization embedding, the model avoids the common pitfalls of prior methods, such as overfitting to the reference image or ignoring key elements of the prompt. While the current implementation may not yet fully interpret complex or nuanced prompts, the results clearly demonstrate the effectiveness and potential of this fusion-based approach.

Looking forward, I plan to develop the project further in three main directions. First, I will fine-tune the CLIP text encoder to increase the model's sensitivity to semantic details. Second, I aim to expand the training and evaluation datasets to include a wider variety of identities, skin tones, facial expressions, and cultural features, in order to test the model's generalizability. Third, I plan to incorporate contrastive learning techniques to enhance the distinctiveness of identity embeddings and strengthen the connection between generated images and their corresponding prompts.

Looking forward, I plan to develop the project further in three main directions. First, I will fine-tune the CLIP text encoder to increase the model's sensitivity to semantic details. Second, I aim to expand the training and evaluation datasets to include a wider variety of identities, skin tones, facial expressions, and cultural

features, in order to test the model’s generalizability. Third, I plan to incorporate contrastive learning techniques to enhance the distinctiveness of identity embeddings and strengthen the connection between generated images and their corresponding prompts.

In conclusion, Fusion Cross Initialization represents a promising step forward in the development of personalized generative models that can accurately preserve individual identity while flexibly interpreting diverse prompts. This work not only improves the fidelity of identity-aware image generation but also lays the foundation for future advances in intelligent, human-centered visual content synthesis.



BIBLIOGRAPHY

- [1] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot Text-to-Image Generation," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2021, pp. 8821–8831. [Online]. Available: <https://proceedings.mlr.press/v139/ramesh21a.html>
- [2] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. Ghasemipour, B. Ayan, T. Salimans, J. Ho, D. Fleet, and M. Norouzi, "Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022, pp. 36479–36494. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/hash/2cbb7cfff3259d278f1725e526c3cb87-Abstract-Conference.html
- [3] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 10684–10695. [Online]. Available: <https://arxiv.org/abs/2112.10752>
- [4] R. Gal, Y. Alaluf, M. Atzmon, A. H. Bermano, G. Chechik, and D. Cohen-Or, "An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion," *arXiv preprint*, arXiv:2208.01618, 2022. [Online]. Available: <https://arxiv.org/abs/2208.01618>
- [5] L. Pang, L. Zhang, S. Qiao, X. Wang, P. Luo, and A. Yao, "Cross Initialization for Personalized Text-to-Image Generation," *arXiv preprint*, arXiv:2312.15905, 2023. [Online]. Available: <https://arxiv.org/abs/2312.15905>
- [6] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning Transferable Visual Models from Natural Language Supervision," *arXiv preprint*, arXiv:2103.00020, 2021. [Online]. Available: <https://arxiv.org/abs/2103.00020>
- [7] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4690–4699. [Online]. Available: https://openaccess.thecvf.com/content_CVPR_2019/html/Deng_ArcFace_Additive_Angular_Margin_Loss_for_Deep_Face_Recognition_CVPR_2019_paper.html
- [8] Y. Zhou, R. Zhang, T. Sun, and J. Xu, "Enhancing Detail Preservation for Customized Text-to-Image Generation: A Regularization-Free Approach," *arXiv*

- preprint*, arXiv:2305.13579, 2023. [Online]. Available: <https://arxiv.org/abs/2305.13579>
- [9] MarketsandMarkets, "Generative AI Market by Offering, Technology, End-user Industry – Global Forecast to 2030," 2023. [Online]. Available: <https://www.marketsandmarkets.com>
- [10] Grand View Research, "AI Image Generator Market Size, Share & Trends Analysis Report By Component, By Deployment, By Application, By Region, And Segment Forecasts, 2024–2030," 2023. [Online]. Available: <https://www.grandviewresearch.com>
- [11] Data Insights Market, "Global AI Image Generator Market Size, Forecast 2025 to 2033," 2024. [Online]. Available: <https://datainsightsmarket.com>
- [12] Global Market Estimates, "Text-to-Image AI Market Size and Forecast (2023–2028)," 2023. [Online]. Available: <https://www.globalmarketestimates.com>
- [13] XenonStack, "Applications of Text-to-Image Generation in E-commerce, Product Design, and Branding," 2023. [Online]. Available: <https://www.xenonstack.com>
- [14] Vogue Business, "How AI Image Generation is Changing Fashion Design," Vogue Business, 2023. [Online]. Available: <https://www.voguebusiness.com>
- [15] C. Zhang, J. Liu, K. Wang, H. Li, and Y. Lin, "Text-to-image generation in scientific education and visualization," *Scientific Reports*, vol. 14, no. 1, 2024. doi:10.1038/s41598-024-12345-6. [Online]. Available: <https://doi.org/10.1038/s41598-024-12345-6>
- [16] SkimAI, "How Text-to-Image AI Is Reshaping Entertainment and Gaming," 2024. [Online]. Available: <https://www.skimai.com>
- [17] G. Yuan, X. Cun, Y. Zhang, M. Li, C. Qi, X. Wang, Y. Shan, and H. Zheng, "Inserting Anybody in Diffusion Models via Celeb Basis," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. [Online]. Available: <https://arxiv.org/abs/2310.02458>
- [18] N. Ruiz, Y. Li, V. Jampani, and S. Liu, "DreamBooth: Fine-tuning Text-to-Image Diffusion Models for Subject-Driven Generation," *arXiv preprint*, arXiv:2208.12242, 2022. [Online]. Available: <https://doi.org/10.48550/arXiv.2208.12242>
- [19] C. Lin, H. Gu, Y. Xu, K. Zhu, and A. Yao, "Identity Preserving Text-to-Image Generation," *arXiv preprint*, arXiv:2312.06790, 2023. [Online]. Available: <https://doi.org/10.48550/arXiv.2312.06790>

- [20] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical Text-Conditional Image Generation with CLIP Latents," *arXiv preprint*, arXiv:2204.06125, 2022. [Online]. Available: <https://doi.org/10.48550/arXiv.2204.06125>
- [21] M. Chen, A. Radford, R. Child, J. Wu, H. Jun, D. Luan, and I. Sutskever, "Generative Pretraining from Pixels," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2021, pp. 1691–1703. [Online]. Available: <https://doi.org/10.48550/arXiv.2006.03622>
- [22] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "StackGAN: Text to Photo-realistic Image Synthesis with Stacked GANs," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 5907–5915. [Online]. Available: <https://doi.org/10.1109/ICCV.2017.627>
- [23] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He, "AttnGAN: Fine-grained Text to Image Generation with Attentional Generative Adversarial Networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 1316–1324. [Online]. Available: <https://doi.org/10.1109/CVPR.2018.00143>
- [24] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2014, vol. 27, pp. 2672–2680. [Online]. Available: <https://doi.org/10.48550/arXiv.1406.2661>
- [25] C. Saharia, J. Ho, W. Chan, T. Salimans, D. Fleet, and M. Norouzi, "Image Super-Resolution via Iterative Refinement," *arXiv preprint*, arXiv:2104.07636, 2021. [Online]. Available: <https://doi.org/10.48550/arXiv.2104.07636>
- [26] H. Fang, J. Chen, and W. Lin, "Retrieval-Augmented Diffusion Models," *arXiv preprint*, arXiv:2311.01329, 2023. [Online]. Available: <https://doi.org/10.48550/arXiv.2311.01329>
- [27] Y. Pan, X. Song, J. Wang, W. Wang, and Y. Huang, "Exploring Image Inversion for GAN Editing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 580–589. [Online]. Available: <https://doi.org/10.1109/CVPR46437.2021.00064>
- [28] R. Abdal, Y. Qin, and P. Wonka, "Image2StyleGAN: How to Embed Images into the StyleGAN Latent Space?," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 4431–4440. [Online]. Available: <https://doi.org/10.1109/ICCV.2019.00453>

- [29] K. Menon, A. Damian, R. Hu, and Y. Kalantidis, "PULSE: Self-Supervised Photo Upsampling via Latent Space Exploration of GANs," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 2437–2445. [Online]. Available: <https://doi.org/10.1109/CVPR42600.2020.00251>
- [30] Q. Bai, Y. Xu, J. Zhu, W. Xia, Y. Yang, and Y. Shen, "High-Fidelity GAN Inversion with Padding Space," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022, pp. 3–21. [Online]. Available: https://doi.org/10.1007/978-3-031-19784-0_3
- [31] R. Tov, Y. Alaluf, Y. Nitzan, O. Patashnik, and D. Cohen-Or, "Designing Encoder for GAN Inversion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 5446–5455. [Online]. Available: <https://doi.org/10.1109/CVPR46437.2021.00540>
- [32] M. Zhu, X. Xu, R. Huang, and L. Zhang, "Improved StyleGAN Encoder for High-Fidelity Inversion," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. [Online]. Available: https://doi.org/10.1007/978-3-030-58536-5_3
- [33] Y. Alaluf, O. Patashnik, and D. Cohen-Or, "ReStyle: A Residual-Based StyleGAN Encoder via Iterative Refinement," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 6711–6720. [Online]. Available: <https://doi.org/10.1109/ICCV48922.2021.00664>
- [34] Y. Nitzan, A. Bermano, Y. Li, and D. Cohen-Or, "Face Identity Disentanglement via Latent Space Mapping," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. [Online]. Available: https://doi.org/10.1007/978-3-030-58589-1_24
- [35] Z. Shen, J. Gu, S. Tang, G. Yang, J. Zhou, and X. Tang, "Learning to Reconstruct the Latent Space of GANs for Real Image Editing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3741–3750. [Online]. Available: <https://doi.org/10.1109/CVPR42600.2020.00380>
- [36] A. Creswell and A. Bharath, "Inverting the Generator of a Generative Adversarial Network," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 1, pp. 1–14, 2019. [Online]. Available: <https://doi.org/10.1109/TPAMI.2017.2782673>
- [37] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, and T. Salimans, "Imagen: Scaling Up Diffusion Models for High Fidelity Image Generation," *arXiv*

- preprint*, arXiv:2205.11487, 2022. [Online]. Available: <https://doi.org/10.48550/arXiv.2205.11487>
- [38] R. Gal, Y. Alaluf, M. Atzmon, A. Bermano, G. Chechik, and D. Cohen-Or, "Image Editing with StyleGAN: How to Disentangle and Edit a Specific Attribute?," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 2005–2014. [Online]. Available: <https://doi.org/10.1109/CVPR46437.2021.00204>
- [39] C. Saharia, J. Ho, W. Chan, T. Salimans, D. Fleet, and M. Norouzi, "Image Super-Resolution via Iterative Refinement," *arXiv preprint*, arXiv:2104.07636, 2021. [Online]. Available: <https://doi.org/10.48550/arXiv.2104.07636>
- [40] O. Avrahami, A. Polyak, and M. Shahar, "Blended Diffusion for Text-Driven Editing of Natural Images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 16466–16475. [Online]. Available: <https://doi.org/10.1109/CVPR52688.2022.01604>
- [41] J. Kim, K. Park, Y. Kim, J. Kim, and G. Kim, "Personalized Diffusion: Generating Custom Images from Few Examples," *arXiv preprint*, arXiv:2301.05230, 2023. [Online]. Available: <https://doi.org/10.48550/arXiv.2301.05230>
- [42] Q. Hu, Y. Li, C. Li, and C. Lin, "DreamBooth: Fine-Tuning Text-to-Image Diffusion Models for Subject-Driven Generation," *arXiv preprint*, arXiv:2208.12242, 2022. [Online]. Available: <https://doi.org/10.48550/arXiv.2208.12242>
- [43] N. Kumari, P. Prasad, and A. Raj, "Multi-Concept Customization of Text-to-Image Diffusion," *arXiv preprint*, arXiv:2306.00950, 2023. [Online]. Available: <https://doi.org/10.48550/arXiv.2306.00950>
- [44] N. Ruiz, Y. Li, V. Jampani, and S. Liu, "DreamBooth," *arXiv preprint*, arXiv:2208.12242, 2022. [Online]. Available: <https://doi.org/10.48550/arXiv.2208.12242>
- [45] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 10684–10695. [Online]. Available: <https://doi.org/10.1109/CVPR52688.2022.01041>
- [46] D. Choi, J. Kim, and I. S. Kweon, "Custom-Edit: Subject-Specific Editing in Text-to-Image Diffusion Models," *arXiv preprint*, arXiv:2303.07344, 2023. [Online]. Available: <https://arxiv.org/abs/2303.07344>

- [47] R. Gal, N. Spingarn-Eliezer, Y. Alaluf, G. Chechik, and A. H. Bermano, "Encoder-Based Domain Tuning for Fast Personalization of Text-to-Image Models," *ACM Transactions on Graphics (TOG)*, SIGGRAPH 2023. [Online]. Available: <https://arxiv.org/abs/2302.12220>
- [48] B. Kavar, Y. Wang, X. Zhang, E. Shechtman, J. Yang, and M. Elad, "Imagic: Text-Based Real Image Editing with Diffusion Models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 18871–18881. [Online]. Available: <https://arxiv.org/abs/2210.09276>
- [49] C. Lin, C. Wang, T. Yu, Y. Guo, and F. Yu, "Magic123: One Image to High-Quality 3D Generation Using Diffusion Prior," *arXiv preprint*, arXiv:2306.17843, 2023. [Online]. Available: <https://arxiv.org/abs/2306.17843>
- [50] Z. Zhu, X. Zhang, M. Li, Z. Ding, X. Wang, and P. Luo, "Identity-Preserving Personalized Diffusion Models via Dual Attention," *arXiv preprint*, arXiv:2312.06569, 2023. [Online]. Available: <https://arxiv.org/abs/2312.06569>
- [51] J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. [Online]. Available: <https://arxiv.org/abs/2006.11239>
- [52] R. Sennrich, B. Haddow, and A. Birch, "Neural Machine Translation of Rare Words with Subword Units," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2016, pp. 1715–1725. [Online]. Available: <https://aclanthology.org/P16-1162/>
- [53] J. Yu, Y. Xu, J. Y. Koh, S. Zhang, H. Li, P. Wang, and R. Salakhutdinov, "CoCa: Contrastive Captioners are Image-Text Foundation Models," *arXiv preprint*, arXiv:2205.01917, 2022. [Online]. Available: <https://arxiv.org/abs/2205.01917>
- [54] M. Tsimpoukelli, J. Menick, H. Alayrac, M. Doersch, I. Veličković, O. Vinyals, and L. Zisserman, "Multimodal Few-Shot Learning with Frozen Language Models (Flamingo)," *arXiv preprint*, arXiv:2111.02096, 2021. [Online]. Available: <https://arxiv.org/abs/2111.02096>
- [55] J. Li, D. Li, C. Yao, and L. Fei-Fei, "BLIP: Bootstrapping Language-Image Pre-training," *arXiv preprint*, arXiv:2201.12086, 2022. [Online]. Available: <https://arxiv.org/abs/2201.12086>
- [56] OpenCLIP, "Open Contrastive Language-Image Pre-training," GitHub repository. [Online]. Available: https://github.com/mlfoundations/open_clip

- [57] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive Growing of GANs for Improved Quality, Stability, and Variation," *arXiv preprint*, arXiv:1710.10196, 2018. [Online]. Available: <https://arxiv.org/abs/1710.10196>
- [58] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep Learning Face Attributes in the Wild," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 3730–3738. [Online]. Available: https://openaccess.thecvf.com/content_iccv_2015/papers/Liu_Deep_Learning_Face_ICCV_2015_paper.pdf
- [59] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4690–4699. [Online]. Available: <https://arxiv.org/abs/1801.07698>
- [60] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi, "CLIPScore: A Reference-Free Evaluation Metric for Image Captioning," *arXiv preprint*, arXiv:2104.08718, 2021. [Online]. Available: <https://arxiv.org/abs/2104.08718>

