Enhancing Identity Preservation in CodeFormer via Identity-Conditioned Codebook Lookup and Feature Modulation

CS4681 - Advanced Machine Learning

Progress Report

Sanjana K. Y. C.

210572P

1.0 Foundational Analysis and Literature Review

The field of blind face restoration (BFR) has made significant strides; however, preserving individual identity in the face of severe degradation remains a formidable challenge. The CodeFormer model represents a notable advancement by reframing BFR as a code prediction task, thereby achieving superior robustness[1]. However, this robustness is achieved through a representational trade-off that can compromise fidelity to identity. This section first deconstructs the CodeFormer paradigm to identify the architectural and conceptual origins of this limitation. Subsequently, it provides a comprehensive review of state-of-the-art methodologies in identity-preserving face restoration, establishing the context and motivation for the proposed enhancements.

1.1 The CodeFormer Paradigm: A Duality of Robustness and Representational Limitation

The central innovation of the CodeFormer framework is its departure from direct pixel-space or continuous latent-space restoration. Instead, it casts BFR as a code prediction task within a discrete, learned proxy space[1]. This paradigm is built upon a vector-quantized autoencoder (VQ-VAE) trained on high-quality face images. The VQ-VAE's encoder and decoder are accompanied by a finite codebook of "visual atoms"—high-quality feature vectors that represent fundamental components of facial structure and texture. During restoration, a low-quality (LQ) input is mapped to a sequence of these codes, which are then passed to the fixed, pre-trained decoder to synthesize the restored image. This approach confers remarkable robustness against severe and unknown degradations. By mapping a corrupted, high-dimensional input to a constrained, low-dimensional sequence of clean codes, the model effectively discards noise and ambiguity, reducing the ill-posed nature of the problem[1].

A key component of this framework is the Transformer-based prediction network. Recognizing that local features in severely degraded images are unreliable for codebook lookup via nearest-neighbor matching, CodeFormer employs a Transformer to model the global composition and long-range dependencies of the input face. This allows the model to infer the most probable code sequence even when local details are entirely lost, ensuring a coherent and plausible facial structure in the output[1].

Despite these strengths, the model's architecture presents an inherent limitation, as acknowledged in the original paper: identity inconsistency[1]. This failure is particularly evident in cases involving "rare visual parts such as accessories" or less common poses like "side faces"[1]. The very mechanism that endows CodeFormer with its robustness—the reliance on a finite, pre-trained codebook—is also the source of its identity preservation weakness. The model's representational capacity is fundamentally constrained by the visual concepts encapsulated within its 1024 codebook vectors. The process of mapping a continuous and corrupted input space to a discrete and clean latent space is necessarily lossy. While this is designed to discard degradation, it can also discard subtle, low-energy signals that define a unique identity, especially if those features are statistically rare within the FFHQ training dataset. If a specific facial attribute, such as a distinctive mole, an atypical eye shape, or the unique structure of a face in profile, is not well-represented by the available combinations of these visual atoms, the model is forced to generate the closest possible approximation using its existing "vocabulary." This results in a regression to the mean, producing a face that is high-quality and plausible but has drifted from the original identity. The Transformer, despite its powerful global modeling capabilities, cannot predict a sequence of codes that do not exist or are not suitable for faithfully reconstructing the target's unique features. This reveals a fundamental tension within the CodeFormer architecture: its strength in generalization and robustness is achieved at the direct expense of its fidelity to outlier identities and fine-grained facial details. Therefore, any meaningful improvement to its identity preservation capabilities must focus not on replacing this robust mechanism, but on providing it with more explicit, identity-aware guidance.

1.2 State-of-the-Art in Identity-Preserving Face Restoration: A Paradigm Shift

To address the challenge of identity preservation, the research community has increasingly moved away from purely blind restoration methods toward guided and conditioned generative models. This evolution is marked by key advancements in loss functions, the use of reference imagery, and novel architectural designs.

A foundational development has been the integration of identity-preserving loss functions directly into the training process. These losses leverage pre-trained, high-performance face recognition networks to provide a strong supervisory signal. The ArcFace loss, an additive angular margin loss, has become a cornerstone of this approach.[2] ArcFace optimizes for a small geodesic distance on the feature hypersphere between the embeddings of two images of the same identity, thereby enhancing intra-class compactness and inter-class discrepancy.[2] By calculating the cosine distance between the ArcFace embeddings of the restored output and the ground-truth image, models can be explicitly penalized for any deviation in identity. Frameworks such as GFP-GAN successfully incorporated this loss to achieve a better balance between perceptual quality (realness) and identity preservation (fidelity).[4] More recent diffusion-based models also employ this strategy, using the gradient of the ArcFace loss to guide the denoising process toward an identity-consistent output.[5]

A more profound paradigm shift has been the emergence of reference-guided restoration. Instead of operating blindly, these methods utilize one or more high-quality reference images of the target subject to provide an unambiguous source of identity information[7]. This transforms the problem from a general restoration task to a personalized one, significantly constraining the solution space and mitigating identity drift[9]. This approach has proven highly effective, as the reference image provides rich, high-fidelity details that may be absent in the degraded input.

To effectively leverage this reference information, novel architectural components have been developed. A prominent innovation is the use of dedicated identity encoders. Models like FaceMe employ a sophisticated encoder that combines features from a general-purpose vision model (e.g., CLIP) with those from a specialized face recognition network (e.g., ArcFace) to extract a robust and disentangled identity vector[8]. This vector captures the essence of the subject's identity, independent of pose, expression, or illumination.

Once a high-fidelity identity vector is extracted, it must be injected into the generative network to guide the synthesis process. This is typically accomplished through conditional feature modulation techniques. These methods allow the identity vector to influence the style and characteristics of the generated output at various layers of the network. Examples include the Channel-Split Spatial Feature Transform (CS-SFT) used in GFP-GAN, which modulates features to balance identity information with restored textures [4], and Adaptive Instance Normalization (AdaIN), which transfers stylistic information by aligning channel-wise feature statistics[10]. In Transformer and diffusion models, this conditioning is often achieved by augmenting self-attention layers with cross-attention mechanisms, where the image features attend to the identity vector, thereby infusing the generation process with identity-specific guidance[10].

The following table provides a comparative analysis of these state-of-the-art approaches, contextualizing the CodeFormer baseline and highlighting the opportunity for the proposed enhancements.

Method	Backbone	Identity Preservation Mechanism	Priors Used	Known Limitations
GFP-GAN[4]	GAN (U-Net + StyleGAN)	Identity Preserving Loss (ArcFace); CS-SFT Layers	Pre-trained StyleGAN2	High reliance on skip connections can introduce artifacts from severely degraded inputs.
CodeFormer (Baseline)[1]	VQ-VAE + Transformer	Global context modeling for code prediction.	Learned Discrete Codebook	No explicit identity loss; struggles with features/poses underrepresented in the codebook.
FaceMe[8]	Diffusion Model	Reference-based Identity Encoder (CLIP+ArcFace) guides cross-attention.	Pre-trained Diffusion Model	Requires high-quality reference images; involves a complex two-stage training process.

Diffuse and Restore[5]	Diffusion Model	Identity Preserving Conditioner (IPC) network; Region-adaptive gradient guidance.	Pre-trained Diffusion Model; ArcFace	Complex guidance mechanism; potential trade-off between identity and overall sharpness.
Proposed Model	VQ-VAE + Transformer	**1.** ArcFace Identity Loss **2.** Identity-Conditioned Transformer **3.** Identity-Aware Feature Modulation	Learned Codebook; ArcFace	To be determined through experimentation.

2.0 Proposed Methodological Enhancements

To systematically address the identified limitations in CodeFormer's identity preservation capabilities, a three-pronged methodological enhancement is proposed. This strategy involves a progressive integration of identity-aware mechanisms, starting with a proven loss-based supervision and advancing to more deeply integrated architectural modifications. These changes are designed to provide the model with explicit identity information at every critical stage of the restoration process, from feature encoding to code prediction and final image synthesis.

2.1 Architectural Modification 1: Integration of an ArcFace-based Identity Preservation Loss

The most direct and empirically validated approach to improving identity preservation is to incorporate an explicit penalty for identity deviation into the model's training objective. The baseline CodeFormer framework relies on a combination of L1 reconstruction loss, perceptual loss, and adversarial loss, none of which directly optimize for identity similarity[1]. This omission allows the model to find solutions that are perceptually plausible but may not correspond to the correct identity.

To rectify this, an identity-preserving loss term, denoted as Lids, will be introduced into the training objectives for Stage II (Transformer Learning) and Stage III (Controllable Feature Transformation Tuning). This loss will be computed using a pre-trained and frozen ArcFace network, a state-of-the-art face recognition model known for its highly discriminative feature embeddings[2]. The loss will be defined as the cosine distance between the ArcFace embeddings of the restored output image,

I_{res}, and the high-quality ground-truth image I_h:

$$L_{ids} = 1 - cos(ArcFace(I_{res}), ArcFace(I_h))$$

This formulation directly measures the angular separation between the identity vectors of the restored and ground-truth faces in the feature hypersphere. A smaller distance corresponds to higher identity similarity. The L_{ids} term will be added as a weighted component to the existing loss functions. For Stage II, the objective becomes $L'_{tf} = L_{tf} + \lambda_{ids} . L_{ids}$, and for Stage III, it will be added to the complete loss. The hyperparameter λ_{ids} will control the relative importance of identity preservation versus reconstruction quality and perceptual realism. This integration will compel the finetuned encoder (E_L) and the Controllable Feature Transformation (CFT) module to generate features that not only reconstruct the image accurately but also maintain the subject's identity as quantified by a powerful recognition model. This strategy has demonstrated significant success in numerous other generative restoration frameworks[4].

2.2 Architectural Modification 2: Identity-Conditioned Codebook Lookup Transformer

While an identity loss provides a powerful supervisory signal, a more fundamental enhancement involves architecturally integrating identity information into the core decision-making process of the model. In CodeFormer, the most critical decision is the Transformer's prediction of the discrete code sequence that represents the restored face. The current implementation performs this prediction based solely on the features extracted from the degraded input, denoted as Z_l. This can be modeled as an unconditional probability: $P(sequence|Z_l)$. This approach is inherently vulnerable when the identity-defining cues within Z_l are weak, ambiguous, or corrupted.

The proposed modification reframes this process as a conditional prediction task $P(sequence|Z_l, V_{id})$, where V_{id} a high-fidelity identity vector is extracted from a reliable source. Providing the Transformer with this explicit and unambiguous identity vector fundamentally constrains the vast search space of possible code combinations. Instead of searching for any plausible face, the model is guided to find the specific sequence of codes that best represents the target identity within the given spatial structure. The Transformer's attention mechanism is uniquely suited for this conditioning. The self-attention layers model the internal spatial relationships within the degraded feature map Z_l, answering the question of "what spatial structure needs to be filled?" By introducing a cross-attention mechanism, the model can then correlate these spatial queries with the identity vector V_{id}, which answers the question of "who is this person?" This forces the Transformer to learn the mapping between a subject's identity and the specific combination of visual atoms required to render their face, leading to a much more informed and identity-consistent code prediction.

The implementation will involve two key components:

1. **Identity Encoder Network:** A separate, lightweight Identity Encoder will be introduced. This network will be built upon a pre-trained and frozen ArcFace model to ensure the extraction of robust, discriminative identity features[8]. During training, it will take a high-quality reference image as input to generate a 512-dimensional identity embedding V_{id}. For inference, it can use either a provided reference image or, in a truly blind scenario, the low-quality input itself, leveraging ArcFace's robustness. A small multi-layer perceptron (MLP) will then project this embedding to match the internal feature dimension of the Transformer module.
2. **Conditioned Transformer Module:** The architecture of the CodeFormer Transformer will be modified. Following each of the existing self-attention blocks, a cross-attention layer will be inserted. In these new layers, the queries (Q) will be derived from the evolving image feature

representation, while the keys (K) and values (V) will be derived from the projected identity vector V_{id}. This ensures that at every stage of processing, the representation is refined and guided by the target identity, making the final code prediction explicitly identity-aware.

2.3 Architectural Modification 3: Identity-Aware Controllable Feature Transformation (IA-CFT)

The final proposed enhancement targets the decoder stage, where the predicted code sequence is translated into the final restored image. The original Controllable Feature Transformation CFT module in CodeFormer modulates the decoder's features (F_d) based on features from the LQ encoder (F_e) to improve fidelity to the input's structure[1]. While this helps preserve pose and expression, it also risks injecting degradation artifacts and does not explicitly leverage identity information. The proposed Identity-Aware CFT (IA-CFT) evolves this module into a dual-pathway system that disentangles structural guidance from identity-specific texture generation.

The decoder's task is twofold: first, to arrange the visual atoms retrieved from the codebook into the correct spatial configuration (e.g., pose, expression, facial layout), and second, to render these atoms with the correct identity-specific style (e.g., skin tone, hair texture, unique facial features). The original CFT conflates these tasks by using a single, potentially corrupted source—the LQ encoder features—for all guidance. The IA-CFT separates these concerns. The structural fidelity path, which preserves the input's pose and composition, will be retained from the original CFT. In parallel, a new identity-styling path will be introduced. This path will leverage the clean identity vector, V_{id}, which contains pure, high-level information about the person's characteristic appearance.

This identity-styling path will be implemented using Adaptive Instance Normalization (AdaIN), a technique proven to be highly effective for style transfer by aligning the channel-wise mean and variance of feature maps[10]. The decoder feature F_d will first be normalized, and then its statistics will be replaced by new scale (γ) and bias (⍰) parameters. These parameters will be predicted by passing the identity vector V_{id} through a dedicated MLP. This mechanism allows the identity vector to directly control the "style" of the decoder's features at multiple scales, effectively instructing the decoder on *how* to render the features, while the original CFT path continues to guide *where* to place them.

The implementation of the IA-CFT module is as follows: The existing CFT operation (Equation 8 in [1]) will be preserved to provide spatial guidance. In parallel, a new branch will take the identity vector V_{id} and pass it through an MLP to predict scale (γ) and bias (⍰) parameters. The decoder feature F_d will then be modulated by an AdaIN operation:

$$AdaIN(F_d, V_{id}) = \gamma(V_{id}) \odot (F_d - \mu(F_d))/\sigma(F_d) + \beta(V_{id})$$

where μ and σ are the channel-wise mean and standard deviation. The outputs of the original CFT path and the new identity-AdaIN path will be additively combined, potentially with learnable or controllable weights, to produce the final modulated decoder feature. This disentangled approach is expected to yield outputs that possess both the structural fidelity of the degraded input and the textural faithfulness of the target identity.

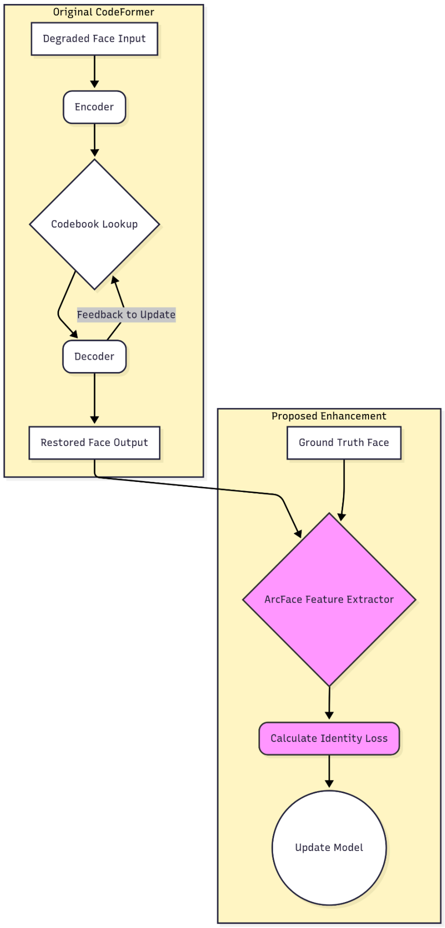

Figure 1: Proposed High-Level Architecture. The original CodeFormer pipeline is enhanced with a dedicated Identity Preservation Module. This module uses a pre-trained ArcFace network to extract identity features from both the ground truth and the restored face, calculating an identity loss that is used to update the generator.

3.0 Experimental Design and Evaluation Protocol

To rigorously validate the efficacy of the proposed methodological enhancements, a comprehensive experimental protocol is designed. This protocol includes careful dataset curation, an adapted training strategy, a multi-faceted evaluation using both quantitative and qualitative metrics, and a series of systematic ablation studies to isolate the contribution of each new component.

3.1 Dataset Curation and Training Strategy

The original CodeFormer was trained on the FFHQ dataset, which, while high-quality, is not ideally suited for developing and evaluating reference-based methods as it generally contains only one image per subject[1]. To facilitate the training of the identity-conditioned components, a new training dataset will be curated from a larger-scale face dataset, such as CelebA-HQ, or by identifying multiple images of the same individuals within FFHQ or similar datasets. This will allow for the formation of training triplets:

$(I_{h-ref}, I_{lq}, I_{h-gt})$, where I_{h-ref} is a high-quality reference image, I_{lq} is a synthetically degraded version of a different image of the same person, and I_{h-gt} is the corresponding high-quality ground truth.

The established three-stage training pipeline of CodeFormer will be adapted to incorporate the proposed modifications:

- **Stage I (Codebook Learning):** This stage will remain unchanged. The goal is to learn a rich and expressive codebook of high-quality visual atoms from a large corpus of faces, as in the original work[1].
- **Stage II (Transformer Learning):** This stage will be significantly modified. The Transformer will be trained with the new cross-attention mechanism, conditioned on identity vectors (V_{id}) extracted from the reference image I_{h-ref}. The training objective will be the sum of the original code-level losses $(L^{token}_{code}\ and\ L^{feat'}_{code})$ and the newly introduced identity preservation loss, Lids, calculated between the restored output and I_{h-gt}.
- **Stage III (IA-CFT Tuning):** The entire network, now including the Identity Encoder, the Conditioned Transformer, and the dual-path IA-CFT module, will be fine-tuned end-to-end. The training will use the complete set of image-level (L1, perceptual, adversarial) and code-level losses, with the L_{ids} term included to ensure consistent identity preservation throughout the full generative process.

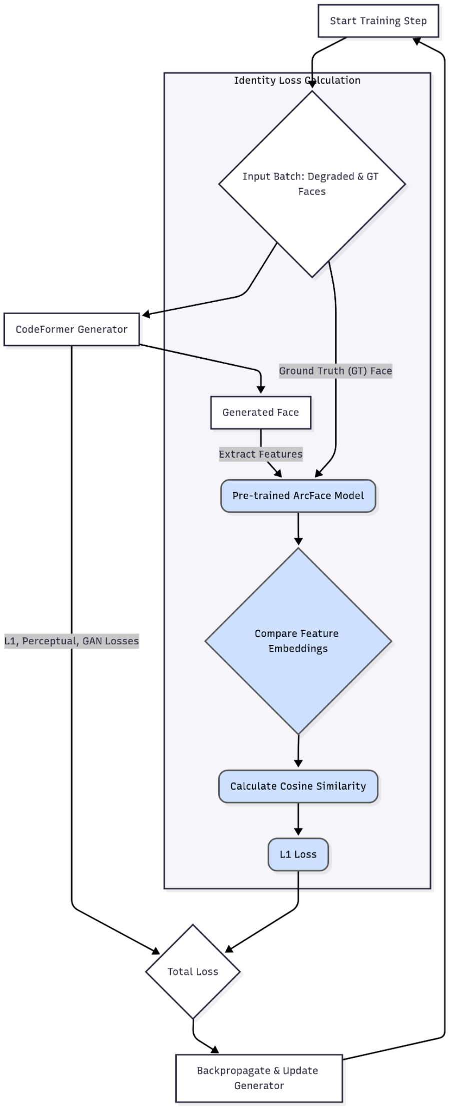
Start Training Step
Identity Loss Calculation
Input Batch: Degraded & GT Faces
CodeFormer Generator
Ground Truth (GT) Face
Generated Face
Extract Features
Pre-trained ArcFace Model
L1, Perceptual, GAN Losses
Compare Feature Embeddings
Calculate Cosine Similarity
L1 Loss
Total Loss
Backpropagate & Update Generator

Figure 2: Detailed Flowchart of the Identity Loss Calculation in a Single Training Step. The generated face and the ground truth face are passed to a frozen ArcFace model to extract high-level feature embeddings. The cosine similarity between these embeddings is then used to compute an L1 loss, which is added to the total loss function to guide the generator towards preserving identity.

3.2 Quantitative and Qualitative Evaluation Metrics

A robust evaluation will be conducted on standard synthetic and real-world test sets, including CelebA-Test, LFW-Test, and the more challenging WIDER-Test[1].

- **Quantitative Metrics:** Performance will be measured using a suite of established metrics to assess different aspects of restoration quality.
  - **Image Quality:** Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and the Learned Perceptual Image Patch Similarity (LPIPS) will be used to evaluate reconstruction accuracy and perceptual quality, following the protocol in[1].
  - **Identity Preservation:** The primary metric for the core objective of this research will be the **IDS score**, calculated as the cosine similarity between the ArcFace embeddings of the restored image and the ground-truth image. Beyond reporting the mean IDS score, a detailed analysis of the score distribution will be performed to assess improvements, particularly in difficult cases where the baseline model fails.
- **Qualitative Metrics:** Visual inspection is critical for evaluating face restoration. To directly probe the model's ability to overcome the specific limitations identified in the original CodeFormer paper, a new challenging test set will be curated. This set will be specifically designed to include:
  - Subjects with prominent and unique accessories (e.g., distinct glasses, hats, earrings) are often normalized or lost by generative models.
  - A significantly higher proportion of non-frontal poses, including profiles and three-quarter views, was used to test the model's robustness to poses underrepresented in the FFHQ training data.
  - A broad range of ethnicities, age groups, and genders to evaluate the fairness and generalizability of the learned codebook and the new identity-conditioning mechanisms.
    Visual comparisons between the baseline CodeFormer and the proposed enhanced model on this challenging dataset will provide clear qualitative evidence of the improvements.

3.3 Ablation Studies

To scientifically dissect the impact of each proposed modification, a rigorous set of ablation studies will be conducted. This is essential for understanding the source of any performance gains and for validating the design choices. A baseline model (the original CodeFormer retrained on the newly curated dataset) will be established, and each new component will be added incrementally. The performance, particularly the LPIPS and IDS scores, will be measured at each step. This systematic approach will provide clear, empirical evidence for the efficacy of the identity loss, the conditioned transformer, and the identity-aware feature modulation module, respectively.

Configuratio n	Identity Loss (Lids)	Conditioned Transformer	IA-CFT Module	LPIPS ↓	IDS ↑
A (Baseline)					
B	✓				
C	✓	✓			
D (Full Model)	✓	✓	✓		

4.0 Projected Timeline and Research Milestones

The proposed research will be executed over a 14-week period, with the timeline structured to ensure systematic progress from foundational work to final evaluation and dissemination.

- **Weeks 1-2: Literature Review and Finalized Methodology:** This initial phase will involve a deep dive into the most recent literature (CVPR, ECCV, NeurIPS 2023-2024) on reference-based restoration, identity-conditioned diffusion models, and advanced feature modulation techniques. The precise architectures of the Identity Encoder, the Conditioned Transformer, and the IA-CFT module will be finalized based on this review.
- **Weeks 3-4: Data Curation and Codebase Preparation:** The primary task will be the curation of the reference-based training, validation, and testing datasets. Concurrently, the official CodeFormer codebase will be forked, and the foundational software scaffolding for the new modules, loss functions, and data loaders will be implemented.
- **Weeks 5-8: Implementation and Training (Stage II):** The core architectural components—the Identity Encoder and the Conditioned Transformer—will be implemented and integrated into the CodeFormer framework. The Stage II model will be trained, with a focus on monitoring the convergence of both the code prediction accuracy and the new identity preservation loss (Lids).
- **Week 9: Mid-Project Evaluation:** A critical checkpoint will be the comprehensive evaluation of the trained Stage II model. Quantitative and qualitative results will be analyzed to verify that the core identity-conditioning mechanism is functioning as expected and providing a tangible improvement in identity preservation before proceeding to the final stage.
- **Weeks 10-11: Implementation and Training (Stage III):** The dual-path IA-CFT module will be implemented and integrated into the decoder. The complete, end-to-end model will then undergo Stage III fine-tuning, optimizing all components jointly.
- **Week 12: Comprehensive Evaluation and Ablation Studies:** The final model will be subjected to the full suite of quantitative and qualitative evaluations across all test sets, including the newly curated challenging set. The ablation studies outlined in Section 3.3 will be executed to isolate the contributions of each new component.
- **Weeks 13-14: Results Analysis and Report/Paper Drafting:** The final phase will be dedicated to a thorough analysis of all experimental results. Visualizations, comparison tables, and statistical analyses will be generated. These findings will be synthesized into the final research report and a draft of a conference-quality paper detailing the methodology, results, and contributions of the work.

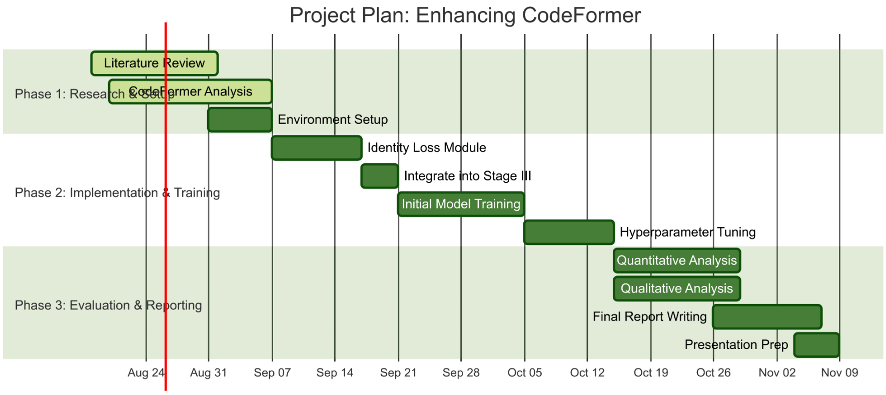

Figure 3: Project Gantt Chart. This chart outlines the key phases, tasks, and timeline for the project. It spans from the initial research and setup phase through implementation, training, and final evaluation, concluding in mid-November 2025.

5.0 Expected Contributions and Future Work

This research is poised to make several significant contributions to the field of blind face restoration by directly addressing the critical challenge of identity preservation within a robust and efficient framework.

5.1 Expected Contributions

1. **A Novel Identity-Preserving Restoration Framework (ID-CodeFormer):** The primary outcome will be a new model, tentatively named ID-CodeFormer, that substantially improves upon the identity preservation capabilities of the original CodeFormer. This enhanced model is expected to demonstrate state-of-the-art performance, particularly in challenging real-world scenarios involving subjects with unique facial features, accessories, and non-canonical poses, where the baseline model is known to falter.
2. **A Methodology for Identity-Conditioned Discrete Representation Learning:** This work will introduce and validate a novel methodology for guiding a Transformer-based code prediction network with an external, high-fidelity identity vector. By demonstrating the effectiveness of augmenting self-attention with identity-based cross-attention in the context of a discrete latent space, this research will provide a new and powerful technique for conditional discrete representation learning that could be applicable to a wide range of generative tasks.
3. **A Disentangled Feature Modulation Module for Guided Synthesis:** The proposed Identity-Aware Controllable Feature Transformation (IA-CFT) module represents a conceptual advance in feature modulation. By disentangling structural guidance (from the degraded input) from stylistic and textural guidance (from a clean identity prior), this module offers a more principled and robust method for balancing fidelity and quality in generative restoration. This design pattern could inform the development of more controllable and faithful generative models in the future.

5.2 Future Work

The framework and concepts developed in this research open several promising avenues for future investigation. The most direct extension is to the domain of **video face restoration**. Maintaining temporal identity consistency across frames is a major challenge, and the proposed identity-conditioning mechanisms could be adapted with temporal attention modules to ensure that a subject's identity remains stable and consistent throughout a video sequence. Furthermore, the core concept of **conditioned codebook lookup** is not limited to face restoration. It could be explored for a variety of other conditional image synthesis tasks, such as text-to-image generation, style transfer, or image editing, where the goal is to generate an image that conforms to both a structural input and a high-level conditioning signal. Finally, exploring methods to dynamically expand or adapt the codebook for out-of-distribution identities could further enhance the model's expressiveness and reduce its reliance on a fixed, pre-trained set of visual atoms.

References

[1] S. Zhou, K. C. K. Chan, C. Li, and C. C. Loy, "Towards Robust Blind Face Restoration with Codebook Lookup Transformer," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2022.

[2] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 4690–4699.

[3] P. Saiprakash, "ArcFace loss function for Deep Face Recognition," *Medium*. Accessed: Aug. 24, 2025. [Online]. Available: https://medium.com/@payyavulasaiprakash/arcface-loss-function-for-deep-face-recognition-e1ff5e173b52

[4] X. Wang, Y. Li, H. Zhang, and Y. Shan, "Towards Real-World Blind Face Restoration With Generative Facial Prior," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 10935–10944.

[5] S. Kim, S.-Y. Cao, J.-F. Hu, and W.-S. Zheng, "A Region-Adaptive Diffusion Model for Identity-Preserving Blind Face Restoration," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, 2024, pp. 293–302.

[6] J. Cho, S. Choi, J. Kim, J. Kim, and S.-H. Bae, "CLR-Face: Conditional Latent Refinement for Blind Face Restoration Using Score-Based Diffusion Models," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, 2024.

[7] D. Park *et al.*, "Reference-Guided Identity Preserving Face Restoration," *arXiv preprint arXiv:2505.21905*, 2025.

[8] Z. Wang, Z. Wang, X. Yang, H. Zeng, and H. Ling, "FaceMe: Robust Blind Face Restoration with Personal Identification," in *Proc. AAAI Conf. Artif. Intell.*, 2025.

[9] Z. Cao, Z. Wang, L. Zheng, Y. Liu, and B. Chen, "InstantRestore: Single-Step Personalized Face Restoration with Shared-Image Attention," *arXiv preprint arXiv:2407.13289*, 2024.

[10] R. Wang *et al.*, "Infinite-ID: Identity-preserved Personalization via ID-semantics Decoupling Paradigm," *arXiv preprint arXiv:2403.11781*, 2024.

[11] J. Cho, S. Choi, J. Kim, J. Kim, and S.-H. Bae, "Conditional Latent Refinement for Blind Face Restoration Using Score-Based Diffusion Models," *arXiv preprint arXiv:2402.06106*, 2024.