# ReF-LDM: A Latent Diffusion Model for Reference-based Face Image Restoration

Chi-Wei Hsiao[1] Yu-Lun Liu[2] Cheng-Kun Yang[1] Sheng-Po Kuo[1]
Yucheun Kevin Jou[1] Chia-Ping Chen[1]

[1]MediaTek [2]National Yang Ming Chiao Tung University

## Abstract

While recent works on blind face image restoration have successfully produced impressive high-quality (HQ) images with abundant details from low-quality (LQ) input images, the generated content may not accurately reflect the real appearance of a person. To address this problem, incorporating well-shot personal images as additional reference inputs could be a promising strategy. Inspired by the recent success of the Latent Diffusion Model (LDM), we propose ReF-LDM—an adaptation of LDM designed to generate HQ face images conditioned on one LQ image and multiple HQ reference images. Our model integrates an effective and efficient mechanism, CacheKV, to leverage the reference images during the generation process. Additionally, we design a timestep-scaled identity loss, enabling our LDM-based model to focus on learning the discriminating features of human faces. Lastly, we construct FFHQ-Ref, a dataset consisting of 20,405 high-quality (HQ) face images with corresponding reference images, which can serve as both training and evaluation data for reference-based face restoration models.

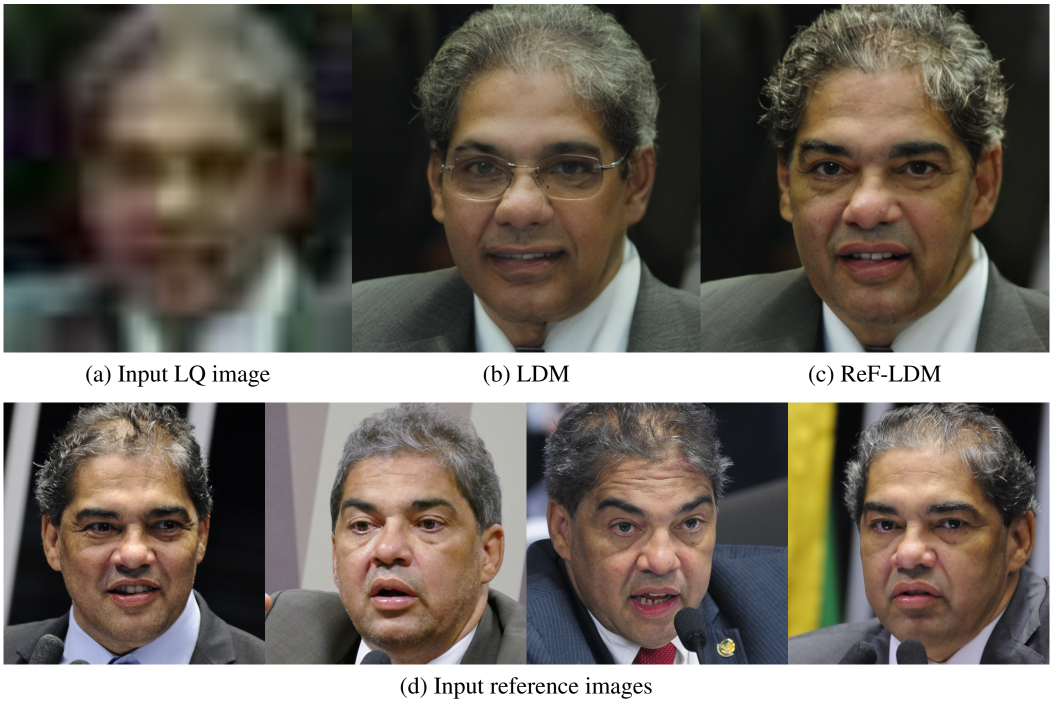
## 1 Introduction

Recent works [5, 26, 32] have achieved impressive results in generating a realistic high-quality (HQ) face image from an input low-quality (LQ) image. However, the important features of a person's face may be corrupted in the LQ image, and thus the reconstructed image may look like a different person. To tackle this problem, besides the LQ image, additional HQ images of this person can be used as reference input. Moreover, allowing multiple reference images may lead to better quality because they offer more comprehensive appearance of this person in different conditions, e.g., different poses, expressions, or lighting.

A previous work [16] has explored using multiple reference images for face restoration. Their method, however, depends on a face landmark model to detect facial components (i.e., eyes, nose, and mouse), which may become unreliable when the input LQ image is severely degraded. Besides, latent diffusion model (LDM) [22] has also been used in different image generating tasks with different input conditions, such as low-resolution images, semantic maps, or sketch images [22, 29].

Inspired by the recent success of LDM, we propose **ReF-LDM** for reference-based face image restoration. Unlike previous conditional LDM methods where their input conditions are usually spatially aligned with the target image, the reference images are not aligned with the target HQ image in our case. Therefore, we design a **CacheKV** mechanism, which effectively and efficiently integrates the reference images, albeit with different poses and expressions. Furthermore, we introduce a timestep-scaled identity loss to drive the reconstructed image to look like the same person of the LQ and reference images. Lastly, we also construct a new large-scale dataset of face images with corresponding reference images, which can serve as both training and evaluation datasets for future reference-based face restoration research.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).



(a) Input LQ image (b) LDM (c) ReF-LDM

(d) Input reference images

Figure 1: **Reference-based face image restoration.** Given an input low-quality face image (a), a Latent Diffusion Model (LDM) can reconstruct a high-quality image (b); however, it may not be faithful to the individual's facial identity. To address this problem, we propose ReF-LDM, which restores a high-quality image with faithful details (c) by utilizing additional reference images (d).

With the above components, our ReF-LDM outperforms recent state-of-the-art methods with a significant improvement in face identity similarity. Extensive ablation studies for the proposed CacheKV mechanism and timestep-scaled identity loss are also conducted and reported. The main contributions of this work can be summarized as:

- We propose ReF-LDM, which features an effective and efficient CacheKV mechanism, for restoring an LQ face image using multiple reference images.
- We introduce a timestep-scaled identity loss, which considers the characteristics of diffusion models and helps ReF-LDM better learn the discriminating features of human identities.
- We construct FFHQ-Ref, a dataset comprising 20,406 high-quality face images and their corresponding reference images, to facilitate the advance of reference-based face image restoration.

## 2 Related work

**Face restoration without personal reference images** Numerous studies have been proposed for blind face image restoration [28, 2, 5, 32, 20, 13, 26]. Recent works such as VQFR [5] and CodeFormer [32] have achieved promising results by exploiting VQGAN, while DAEFR [26] further employs a dual-branch encoder to mitigate the domain gap between LQ and HQ images. Inspired by the success of diffusion models, several works [23, 31, 17, 27, 25] have adopted diffusion models for face image restoration. However, as these methods do not leverage reference images, the restored images may differ from the authentic facial appearance of a person, especially when an input image is severely degraded.

**Face restoration with personal reference images** Several methods [14, 15, 16, 19] have attempted to utilize additional reference images to enhance personal fidelity in face restoration. GFRNet [14] warps a single reference image to match the face pose of the LQ image, while ASFNet [15] selects the reference image with the closest matching facial landmarks to serve as the network input. Closer to the setting of this work, DMDNet [16] also utilizes multiple reference images. It detects facial landmarks on the LQ image and the reference images to extract features of facial components, and then integrates these features into the model by querying the corresponding components. However, their method relies on landmark detection, which may not be robust on severely degraded LQ images. In contrast, our ReF-LDM implicitly learns the correspondences between the features of the LQ image and the reference images, without the need for landmark detection. From a different perspective, MyStyle [19] adopts a per-person optimization setting, leveraging hundreds of images of an individual to define a personalized subspace within the latent space of a StyleGAN [12]. In comparison, our approach offers greater flexibility, capable of utilizing one to several reference images without the need for personalized model optimization for each individual.

**Latent diffusion models with image conditions** Previous work demonstrates that LDM can generate an image from a low-resolution image by simple channel-axis concatenation [22]. However, reference images in our task are not spatially aligned with the target HQ image, thus requiring a more sophisticated integration mechanism. MasaCtrl [1] achieves text-to-image synthesis with a single reference image by replacing the original keys and values tokens with those from the reference image. However, their solution requires passing the reference image through the denoising network for multiple timesteps, which increases computation and limits its feasibility for extending to multiple reference images. In contrast, we propose an efficient CacheKV mechanism that leverages multiple reference images by eliminating the redundant network passes.

## 3 The proposed ReF-LDM model

In this section, we present the proposed ReF-LDM model. We introduce the network architecture in Sec. 3.1, where a CacheKV mechanism is designed to leverage reference images. We illustrate how to train our model with the timestep-scaled identity loss in Sec. 3.2.

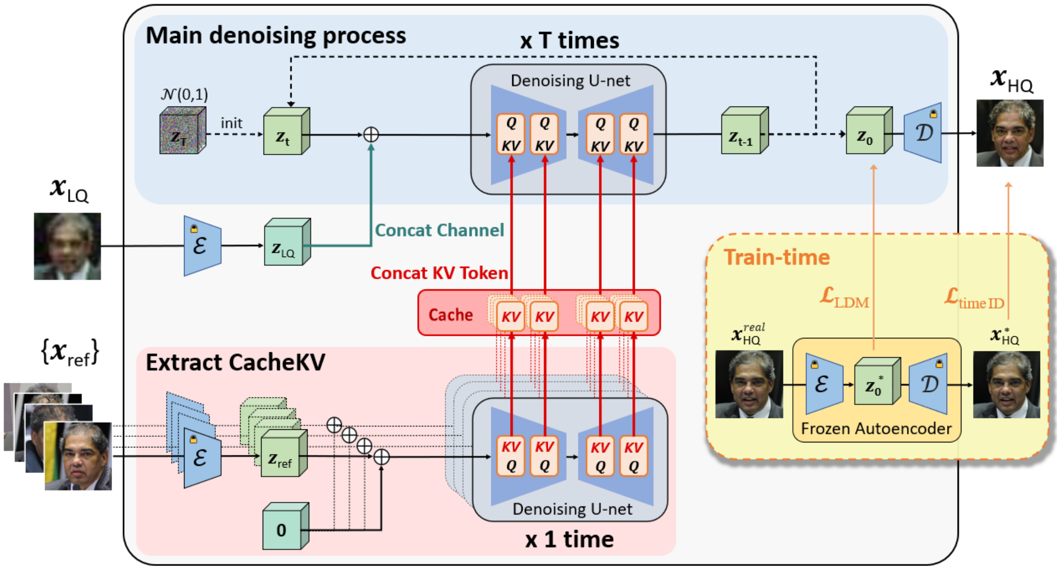


Figure 2: **The proposed ReF-LDM pipeline.** Our model accepts a low-quality image and multiple high-quality reference images as input and generates a high-quality image. The blue top panel alone represents a typical LDM [22] denoising process. For an LQ image $\mathbf{x}_{\text{LQ}}$, we concatenate its latent $\mathbf{z}_{\text{LQ}}$ with $\mathbf{z}_t$ along the channel axis to serve as the input for the denoising U-net. For the reference images $\{\mathbf{x}_{\text{ref}}\}$, we design a **CacheKV** mechanism, depicted in the red panel, to extract and cache their key and value tokens using the same denoising U-net for just one time. These cached KV tokens can then be utlized repeatedly in each of the $T$ timesteps of the main denoising process. During training, we adopt the classic LDM loss ($\mathcal{L}_{\text{LDM}}$) and introduce a timestep-scaled identity loss ($\mathcal{L}_{\text{time ID}}$).

### 3.1 Model architecture of ReF-LDM

The proposed ReF-LDM accepts an input LQ image and multiple reference images to generate a target HQ image. Its model architecture is based on the latent diffusion model [22], with additional designs to incorporate the input LQ image and the reference images.

#### 3.1.1 Preliminaries on Latent Diffusion Model

To generate an image, an image diffusion model [8] starts from a noisy image $\mathbf{x}_T \in \mathbb{R}^{H\times W\times 3}$, initialized with a Gaussian distribution, and gradually denoises it to a clean image $\mathbf{x}_0$ with a denoising network over $T$ timesteps. A latent diffusion model [22] operates similarly, but the diffusion process takes place in a more compact latent space of a pre-trained and frozen autoencoder (encoder $\mathcal{E}$ and decoder $\mathcal{D}$). That is, it begins with a random latent $\mathbf{z}_T \in \mathbb{R}^{H_z\times W_z\times C_z}$ and progressively denoises it to a clean latent $\mathbf{z}_0$. A clean image is then generated by passing the clean latent through the decoder during the inference phase, i.e., $\mathbf{x}_0 = \mathcal{D}(\mathbf{z}_0)$; conversely, a ground truth clean latent is obtained by encoding a clean image with the encoder during the training phase, i.e., $\mathbf{z}_0^* = \mathcal{E}(\mathbf{x}_0^*)$. A typical choice for the denoising network is a U-net with self-attention layers at multiple scales.

#### 3.1.2 CacheKV: a mechanism for incorporating reference images

As illustrated in Fig. 2, our ReF-LDM leverages an input LQ image $\mathbf{x}_{\text{LQ}}$ and multiple reference images $\{\mathbf{x}_{\text{ref}}\}$ to generate a target HQ image $\mathbf{x}_{\text{HQ}}$. For an LQ image, we simply concatenate its latent encoded by the frozen encoder, $\mathbf{z}_{\text{LQ}} = \mathcal{E}(\mathbf{x}_{\text{LQ}})$, with the diffusion denoising latent $\mathbf{z}_t$ along the channel axis to serve as the input of the denoising U-net. For reference images, we design a **CacheKV** mechanism. Essentially, we extract and cache the features of reference images using the same denoising U-net just once; these cached features can then be used repeatedly at each of the $T$ timesteps in the main denoising process. Specifically, we pass the encoded latent of each reference image, $\mathbf{z}_{\text{ref}} = \mathcal{E}(\mathbf{x}_{\text{ref}})$, through the U-net to extract their keys and values (KVs) at each self-attention layer and store them in a CacheKV. Subsequently, during the main diffusion process, within each self-attention layer of the U-net, we concatenate the reference KVs (from the corresponding self-attention layer) with the main KVs along the token axis. This mechanism enables the U-net to incorporate the additional KVs from the reference images into the main denoising process. When extracting KVs from the reference images, we use a timestep embedding of $t = 0$ and pad $\mathbf{z}_{\text{ref}}$ with a zero tensor to accommodate the additional channels introduced for the LQ image.

To summarize, for inference, we first run the U-net once to extract CacheKV from the reference images; subsequently, we proceed through the main denoising process for $T$ timesteps, during which the U-net integrates $\mathbf{z}_{\text{LQ}}$ and reference CacheKV. For training, in each iteration, we first run the U-net to extract CacheKV, and then we run the U-net again to estimate the target latent from a sampled noisy latent $\mathbf{z}_t$, incorporating the conditions $\mathbf{z}_{\text{LQ}}$ and reference CacheKV.

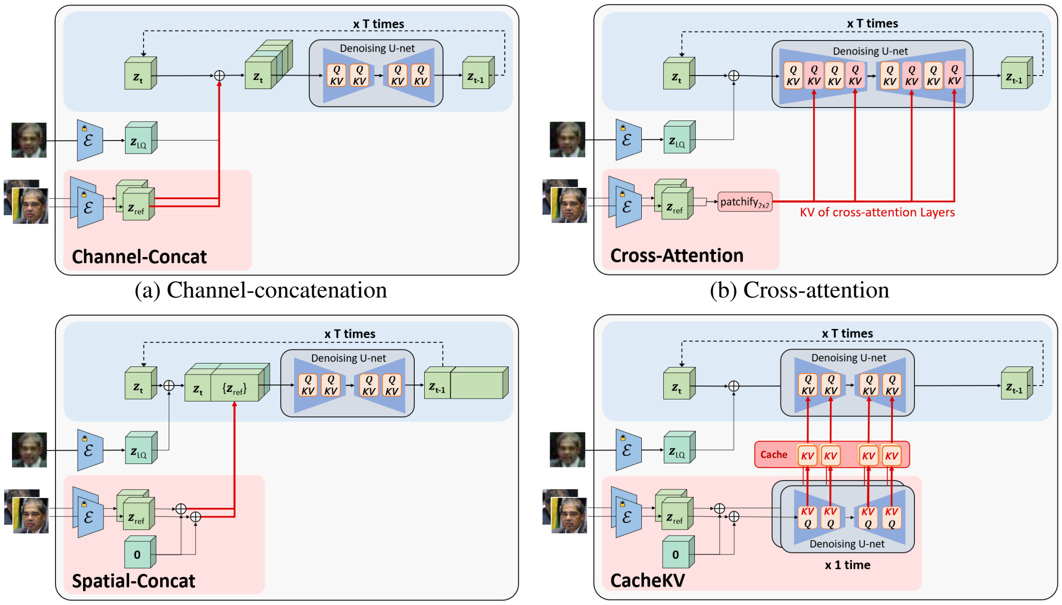
#### 3.1.3 Comparing CacheKV with other designs

There are other intuitive designs for integrating the reference latents $\{\mathbf{z}_{\text{ref}}\}$ into the diffusion denoising process. However, they are either ineffective or computationally inefficient compared to the proposed CacheKV. The quantitative evaluation and computational analysis is reported in Sec. 5.2.1. We depict these designs in Fig. 3 and provide an intuitive explanation as follows:

- **Channel-concatenation**: Concatenating the condition with $\mathbf{z}_t$ along the channel axis works well for LQ images (and for other 2D conditions such as semantic maps [22]); however, it is not effective for reference images. A critical difference between these conditions is that—while the LQ image is spatially aligned with the target HQ image, the reference images are not. Therefore, it is challenging for the model to leverage reference images using simple channel-concatenation.
- **Cross-attention**: Cross-attention layers have been proven useful for text conditions in text-to-image models [22]. In our ablation experiment, we insert a cross-attention layer after each self-attention layer and use the reference latents $\{\mathbf{z}_{\text{ref}}\}$ to produce keys and values. While cross-attention appears to have the potential to address the spatial misalignment problem, it still fails to effectively utilize the reference images. The difference between our CacheKV and the cross-attention setting is that CacheKV provides the reference images in a more aligned feature space for the main denoising process to leverage. Specifically, the CacheKV is extracted using the same U-net and the corresponding self-attention layer as in the main denoising process. In

contrast, the cross-attention setting processes the reference images only with the frozen encoder, resulting in features that are less aligned with those in the U-net of the denoising process.

- **Spatial-concatenation**: Concatenating $\{\mathbf{z}_{\text{ref}}\}$ with $\mathbf{z}_t$ along the spatial dimension to serve as the input for U-net also effectively leverages the reference images. Conceptually, spatial-concatenation treats reference images in a very similar way to our CacheKV. In both mechanisms, $\{\mathbf{z}_{\text{ref}}\}$ are processed through the denoising U-net, allowing the reference KVs to be accessed by the queries (Qs) of the main diffusion latent $\mathbf{z}_t$. However, spatial-concatenation requires significantly more computational resources compared to our CacheKV. It passes $\{\mathbf{z}_{\text{ref}}\}$ with $\mathbf{z}_t$ to the U-net at each of the $T$ denoising timesteps, whereas CacheKV only passes $\{\mathbf{z}_{\text{ref}}\}$ through the U-net once. Moreover, spatial-concatenation also requires significantly more GPU memory, as the spatial size of the input for the U-net increases with the number of reference images. As for a self-attention layer in the U-net, both mechanisms increase memory usage; CacheKV introduces additional reference KVs, while spatial-concatenation introduces reference QKVs.



(a) Channel-concatenation

(b) Cross-attention

(c) Spatial-concatenation

(d) CacheKV (ours)

Figure 3: Different mechanisms for incorporating reference images into the main denoising process.

## 3.2 Timestep-scaled identity loss

### 3.2.1 Timestep-scaled identity loss

As this work aims for face image restoration, we employ the identity loss to enhance face similarity, which is adopted in many face-related tasks [9, 21, 28]. The identity loss minimizes the distance within the embedding space of a face recognition model, thereby capturing the discriminating features of human faces more effectively than the plain RGB pixel space. In our experiments, we use the ArcFace model [3] with cosine distance between the 1D embedding vectors as the identity loss.

However, naively adding identity loss to the training of ReF-LDM significantly worsens the image quality. One possible explanation might be that, the one-step model prediction $\mathbf{x}_{0|\text{t}} = \mathcal{D}(\mathbf{z}_0|\mathbf{z}_t)$ at a very noisy timestep (e.g., $t = T$) is very different from a natural face image and thus out of the distribution that the ArcFace model is trained on; therefore, the identity loss provides ineffective supervision for diffusion models at large timesteps.

Based on this assumption, we propose a timestep-scaled identity loss, where a timestep-dependent scaling factor is introduced to scale down the identity loss when a larger timestep is sampled in a training step. Specifically, the timestep-scaled identity loss is defined as:

$$\mathcal{L}_{\text{time ID}} = \sqrt{\bar{\alpha}_t} \cdot \mathcal{L}_{\text{ID}} = \sqrt{\bar{\alpha}_t} \cdot \left(1 - \frac{R(\mathbf{x}) \cdot R(\mathbf{x}^*)}{\|R(\mathbf{x})\| \|R(\mathbf{x}^*)\|}\right), \quad (1)$$

where $R$ is a face recognition model, and $\sqrt{\bar{\alpha}_t}$ follows the definition in a typical diffusion process [8, 22] in which a noisy latent $\mathbf{z}_t$ is sampled given a clean latent $\mathbf{z}_0^*$ as:

$$q(\mathbf{z}_t|\mathbf{z}_0^*) = \mathcal{N}(\sqrt{\bar{\alpha}_t}\mathbf{z}_0^*, (1-\bar{\alpha}_t)\mathrm{I}) \tag{2}$$

#### 3.2.2 Training ReF-LDM with timestep-scaled identity loss

We train our ReF-LDM with the classic LDM loss and the proposed timestep-scaled identity loss:

$$\mathcal{L}_{total} = \mathcal{L}_{\mathrm{LDM}} + \lambda_{\mathrm{time\,ID}}\,\mathcal{L}_{\mathrm{time\,ID}} \tag{3}$$

Recall that the denoising U-net estimates the target latent in the latent space of the frozen autoencoder, and a typical $\mathcal{L}_{\mathrm{LDM}}$ is computed as the L1 distance between the estimated latent and the target latent. To compute the identity loss with the face recognition model, which accepts an image as input, we decode the estimated latent into the image space using the frozen decoder, i.e, $\mathbf{x}_0 = \mathcal{D}(\mathbf{z}_0)$. The experiments in Sec. 5.2.2 show that timestep-scaled identity loss can improve face similarity without degrading image quality, unlike the naive usage of identity loss.

## 4 FFHQ-Ref dataset

Recent works for non-reference-based face restoration commonly train their models with FFHQ dataset [12], which comprises 70,000 high-quality face images of wide appearance variety with appropriate licenses crawled from Flickr. These images are not provided with reference labels originally; however, we find that a good portion of the images are of the same identities. Thus, we construct a reference-based dataset—FFHQ-Ref—based on the FFHQ dataset, with careful consideration described as follows.

### 4.1 Finding reference images of the same identity

To determine whether two images belong to the same identity, we utilize the face recognition model ArcFace [3]. Specifically, we first extract the 1D ArcFace embeddings for all images. Then, for each image, we compute the cosine distances between its embedding and the embeddings of all other images. A distance less than a threshold $r = 0.4$ indicates that the images are valid references belonging to the same person. Following this procedure, we identify 20,405 images with corresponding reference images.

### 4.2 Splitting data according to identity

To enable the FFHQ-Ref dataset to serve as both training and evaluation datasets for reference-based face restoration models, we divide the images into train, validation, and test splits. However, random data splitting may result in the train and test splits containing images of the same individual, which is not ideal for a fair evaluation. To ensure that all images of a single identity are assigned to only one data split, we group the images based on their identities. Specifically, we consider identity grouping as a graph problem, where each image acts as a vertex and any pair of images with a distance less than $r$ are connected by edges. We then apply the connected component algorithm from graph theory, where each connected component represents a group of images belonging to the same person. Finally, we identified 6,523 identities and divided them into three splits: a train split with 18,816 images of 6,073 identities, a validation split with 732 images of 300 identities , and a test split with 857 images of 150 identities. We report more statistics in Appendix C.

### 4.3 Constructing evaluation dataset with practical considerations

**Practical Considerations** For a fair and meaningful evaluation, the input reference images should not be excessively similar to the target image; hence, we set a minimum cosine distance threshold of 0.1 for the test set. Additionally, we manually check the images in the test split to verify that all reference images indeed correspond to the same identity. Furthermore, in the context of reference-based face restoration applications, it is preferable to select input reference images that capture a more comprehensive representation of a person's appearance, such as varying face poses or expressions. Although a target image in the test split of our FFHQ-Ref may have two to nine reference images, different reference-based methods may have their own constraints on the maximum number of input reference images. To emulate a more representative set of reference images, we sort all available reference images of a target image using farthest point sampling on the ArcFace distance.

**Degradation synthesis for input LQ images** For synthesizing input LQ images from ground truth HQ images, we follow the degradation model used in previous works [28, 5, 32]:

$$\mathbf{x}_{\text{LQ}} = \{[(\mathbf{x}_{\text{HQ}} * k_\sigma) \downarrow_r + n_\delta]_{\text{JPEG}_q}\} \uparrow_r, \tag{4}$$

where an HQ image is blurred with a Gaussian kernel $k_\sigma$, downsampled by $r$ scale, added with a Gaussian noise $n_\delta$, compressed with JPEG quality level $q$, and upscaled to the original size.

We construct two evaluation datasets with different degradation levels:

- FFHQ-Ref-Moderate: $\sigma$, $r$, $\delta$, and $q$ are sampled from $[0, 8]$, $[1, 8]$, $[0, 15]$, and $[60, 100]$.
- FFHQ-Ref-Severe: $\sigma$, $r$, $\delta$, and $q$ are sampled from $[8, 16]$, $[8, 32]$, $[0, 20]$, and $[30, 100]$.

### 4.4 Comparison between FFHQ-Ref and existing datasets

Table 1 summarizes the differences between our proposed FFHQ-Ref and existing datasets. While the CelebRef-HQ dataset [16] has been constructed to train and evaluate reference-based face restoration models, our FFHQ-Ref dataset contains twice as many images and six times the number of identities compared to CelebRef-HQ. Moreover, built upon FFHQ [12], FFHQ-Ref provides superior image quality over CelebRef-HQ, as indicated by the lower NIQE score (3.68 vs. 3.97). Some ground-truth images in CelebRef-HQ are affected by watermarks and mirror padding artifacts, as shown in Appendix B.

Table 1: Comparison between the proposed FFHQ-Ref and existing datasets.

| Dataset | With reference | Licensed | Quality | Images | Identities |
|---|---|---|---|---|---|
| FFHQ [12] | | ✓ | ✓ | 70,000 | - |
| CelebRef-HQ [16] | ✓ | | | 10,555 | 1,005 |
| **FFHQ-Ref** | ✓ | ✓ | ✓ | **20,405** | **6,523** |

## 5 Experiments

In this section, we describe the experimental setup in Sec. 5.1, discuss ablation studies in Sec. 5.2, and provide the comparison between our ReF-LDM and the state-of-the-art methods in Sec. 5.3

### 5.1 Experimental setup

#### 5.1.1 Implementation details

To exploit more ground truth images without available reference images, we use 68,411 images in the FFHQ dataset to train a VQGAN [4] as the frozen autoencoder and an LDM with only LQ condition. We then finetune our ReF-LDM from the LQ-conditioned LDM with the 18,816 images in our FFHQ-Ref dataset. All models are trained excluding the test split images to ensure fair evaluation on our FFHQ-Ref benchmark. In our experiments, we adopt a 512x512 image resolution, fix the number of reference images to five, and set loss scale $\lambda_{\text{time ID}}$ to 0.1. During training, we synthesize input LQ images with $\sigma$, $r$, $\delta$, and $q$ sampled from $[0, 16]$, $[1, 32]$, $[0, 20]$, and $[30, 100]$, respectively. For inference, we use 100 DDIM [24] steps and a classifier-free-guidance [7] with a scale of 1.5 towards reference images. We provides more implementation details in the Appendix G.

#### 5.1.2 Evaluation datasets and metrics

For evaluation datasets, we use the test split of our FFHQ-Ref with two different degradation levels: severe and moderate. In addition, previous non-reference-based methods commonly use CelebA-Test [28] for evaluation, which comprises 3,000 LQ and HQ image pairs sampled from the CelebA-HQ dataset [11]. Therefore, we follow the same procedures described in Sec. 4 to construct a subset of 2,533 images with available reference images, termed CelebA-Test-Ref.

For evaluation metrics, we adopt the identity similarity (IDS) [5, 32], which is the cosine similarity calculated using the face recognition model ArcFace [3]. We also use the widely used perceptual metrics LPIPS [30]. As face pixels are more of concern in the task of face restoration, we also measure the face-region LPIPS (fLPIPS), which is the LPIPS calculated using only the pixels in face regions. For assessing no-reference image quality, we adopt NIQE [18]. Furthermore, we measure the FID [6], using 70,000 images from the FFHQ dataset as the target distribution.

## 5.2 Ablation studies

We provide the ablation studies of the proposed CacheKV, timestep-scaled identity loss, and the number of input reference images. In each ablation experiment, we fine-tune the model for 50,000 steps from the same LDM pre-trained without reference images. We compare the difference settings with the FFHQ-Ref-Severe dataset.

### 5.2.1 CacheKV and other mechanisms

The CacheKV is proposed for integrating the input reference images into the diffusion denoising process. We compare it with other mechanisms illustrated in Sec. 3.1.3. According to Table 2, channel-concatenation and cross-attention fail to leverage reference images to improve the identity similarity (IDS). In contrast, both spatial-concatenation and our CacheKV significantly enhance IDS. Moreover, our CacheKV is more computationally efficient than spatial-concatenation, requiring only 20% of the inference time and 39% of the GPU memory.

Table 2: Comparison between CacheKV and other mechanisms for input reference images (run with five reference images on a single GTX 1080).

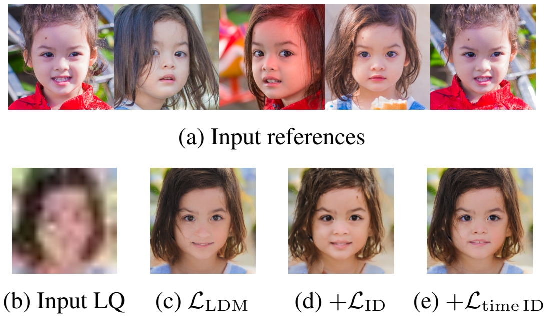
| | IDS↑ | NIQE↓ | LPIPS↓ | Inference time↓ | Memory↓ |
|---|---|---|---|---|---|
| Channel-concatenation | 0.23 | 4.49 | 0.46 | 4.17 | 1.77 |
| Cross-attention | 0.23 | 4.56 | 0.46 | 14.54 | 2.80 |
| Spatial-concatenation | 0.69 | 4.84 | 0.43 | 58.36 | 7.44 |
| **CacheKV** | 0.65 | 4.38 | 0.43 | 12.15 | 2.87 |

Table 3: Ablation results for the timestep-scaled identity loss.

| Loss | IDS↑ | NIQE↓ |
|---|---|---|
| $\mathcal{L}_{\text{LDM}}$ | 0.52 | 4.56 |
| $\mathcal{L}_{\text{LDM}} + \mathcal{L}_{\text{ID}}$ | 0.69 | 6.56 |
| $\mathcal{L}_{\text{LDM}} + \mathcal{L}_{\text{time ID}}$ | 0.65 | 4.38 |

Table 4: Design choices for ID loss scaling.

| Scale for ID loss | IDS↑ | NIQE↓ |
|---|---|---|
| $\sqrt{\bar{\alpha}_t}$ | 0.65 | 4.38 |
| $\mathbf{1}_{t<100}$ | 0.52 | 4.55 |
| $\mathbf{1}_{t<500}$ | 0.61 | 4.44 |


(a) Input references

(b) Input LQ (c) $\mathcal{L}_{\text{LDM}}$ (d) $+\mathcal{L}_{\text{ID}}$ (e) $+\mathcal{L}_{\text{time ID}}$

Figure 4: Visual ablation results for the timestep-scaled identity loss.

### 5.2.2 Timestep-scaled identity loss

To validate the benefits of the proposed timestep-scaled identity loss, we train ReF-LDM with three different loss settings: without identity loss ($\mathcal{L}_{\text{LDM}}$), with naive identity loss ($\mathcal{L}_{\text{LDM}} + \mathcal{L}_{\text{ID}}$), and with the proposed timestep-scaled identity loss ($\mathcal{L}_{\text{LDM}} + \mathcal{L}_{\text{time ID}}$). As show in Table 3 and Fig. 4, while the naive identity loss can improve identity similarity (IDS), our timestep-scaled identity loss can do so without sacrificing the image quality (NIQE).

As explained in Sec. 3.2, we employ $\sqrt{\bar{\alpha}_t}$ to scale down the identity loss for a larger and noisier timestep $t$. In Table 4, we compare this design choice with other alternative scaling factors, $\mathbf{1}_{t<100}$ and $\mathbf{1}_{t<500}$, which apply the identity loss only when the sampled timestep $t$ is smaller than 100 or 500, respectively. The results suggest that $\sqrt{\bar{\alpha}_t}$ is more effective than the alternatives.

### 5.2.3 Multiple input reference images

There are two to nine reference images for a target image in the test split of our FFHQ-Ref. While we fix the number of reference images to five when training ReF-LDM, the proposed CacheKV mechanism has the flexibility to take varying number of reference images during inference. To validate the effectiveness of utilizing multiple reference images, we evaluate ReF-LDM with a maximum of 1, 3, 5, and 8 reference images, respectively. As shown in Table 5, using more reference images significantly improves the identity similarity (from 0.52 to 0.66). However, increasing the number of reference images also increases the computation time, as shown in Table 6. Since using eight reference images encounters an out-of-memory issue on a single GTX 1080, we use at most five reference images in our experiments for simplicity.

Table 5: Image quality with different numbrs of reference images.

| Max num refs | IDS↑ | LPIPS↓ |
|---|---|---|
| 1 | 0.52 | 0.45 |
| 3 | 0.62 | 0.44 |
| 5 | 0.65 | 0.43 |
| 8 | 0.66 | 0.43 |

Table 6: Inference time with different numbers of reference images.

| Num refs | Time@1080↓ | Time@3090↓ |
|---|---|---|
| 1 | 4.86 | 3.00 |
| 3 | 7.09 | 3.54 |
| 5 | 12.03 | 4.51 |
| 8 | out-of-memory | 6.26 |

## 5.3 Comparison with state-of-the-art methods

Table 7: Comparison of ReF-LDM with state-of-the-art methods across three benchmarks. Note the highlighting 1st, 2nd, and a gray cell indicating evaluation data leakage for prior methods .

| | FFHQ-Ref-Severe | | | | FFHQ-Ref-Moderate | | | | CelebA-Test-Ref | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | IDS↑ | fLPIPS↓ | LPIPS↓ | FID↓ | IDS↑ | fLPIPS↓ | LPIPS↓ | FID↓ | IDS↑ | fLPIPS↓ | LPIPS↓ |
| CodeFormer [32] | 0.323 | 0.108 | 0.398 | 51.51 | 0.760 | 0.084 | 0.301 | 38.78 | 0.660 | 0.092 | 0.340 |
| VQFR [5] | 0.308 | 0.112 | 0.415 | 52.96 | 0.659 | 0.089 | 0.324 | 36.77 | 0.558 | 0.096 | 0.352 |
| DAEFR [26] | 0.294 | 0.118 | 0.435 | 49.08 | 0.614 | 0.093 | 0.333 | 33.86 | 0.491 | 0.101 | 0.367 |
| LDM | 0.231 | 0.125 | 0.453 | 34.40 | 0.753 | 0.095 | 0.344 | 32.16 | 0.663 | 0.093 | 0.368 |
| DMDNet [16]$^{\dagger}$ | 0.185 | 0.162 | 0.511 | 72.66 | 0.810 | 0.096 | 0.348 | 36.60 | 0.752 | 0.097 | 0.362 |
| ReF-LDM | 0.676 | 0.110 | 0.429 | 37.60 | 0.840 | 0.088 | 0.332 | 33.05 | 0.779 | 0.093 | 0.368 |

$^{\dagger}$As DMDNet encounters landmark detection failures and fails to yield results for 214/857, 29/857, and 488/2,533 images on the three benchmarks respectively, we compute the metrics for DMDNet using the remaining images.

### 5.3.1 Quantitative comparison

We compare our ReF-LDM with state-of-the-art methods on FFHQ-Ref-Severe, FFHQ-Ref-Moderate, and CelebA-Test-Ref. Table 7 reports the performance of competing methods in terms of IDS, fLPIPS, LPIPS, and FID (targeting the FFHQ image distribution). Without the information in reference images, the existing non-reference-based restoration methods (CodeFormer [32], VQFR [5], and DAEFR [26]) fail to preserve the facial identity, leading to significantly lower IDS. The reference-based method, DMDNet [16], fails to restore the severely degraded images because it depends on unreliable facial landmark detection, reflected by higher fLPIPS. In contrast, our ReF-LDM consistently outperforms DMDNet on identity similarity and other metrics, owing to the proposed CacheKV mechanism and timestep-scaled identity loss, which effectively leverage the input reference images without the need for landmark detection. We also note that our method exhibits slightly inferior results in LPIPS metric. This is due to the difference in the background pixels, we provide further details in the Appendix D. It is also worth mentioning that the competing methods benefit from data leakage on the FFHQ-Ref benchmarks, as their models are trained with the entire FFHQ dataset or with a different train split than the identity-based one in the proposed FFHQ-Ref.

### 5.3.2 Qualitative comparison

In Fig. 5, we present a qualitative comparison between our ReF-LDM, the pre-trained LDM without reference images, CodeFomer (a SOTA non-reference-based method), and DMDNet (a SOTA reference-based method). Given the severely degraded image, DMDNet generates distorted face images based on incorrectly detected landmarks. While CodeFormer yields realistic face images, it does not preserve the facial identity well. In contrast, our ReF-LDM produces results that are both realistic and faithful to the individual's facial identity.

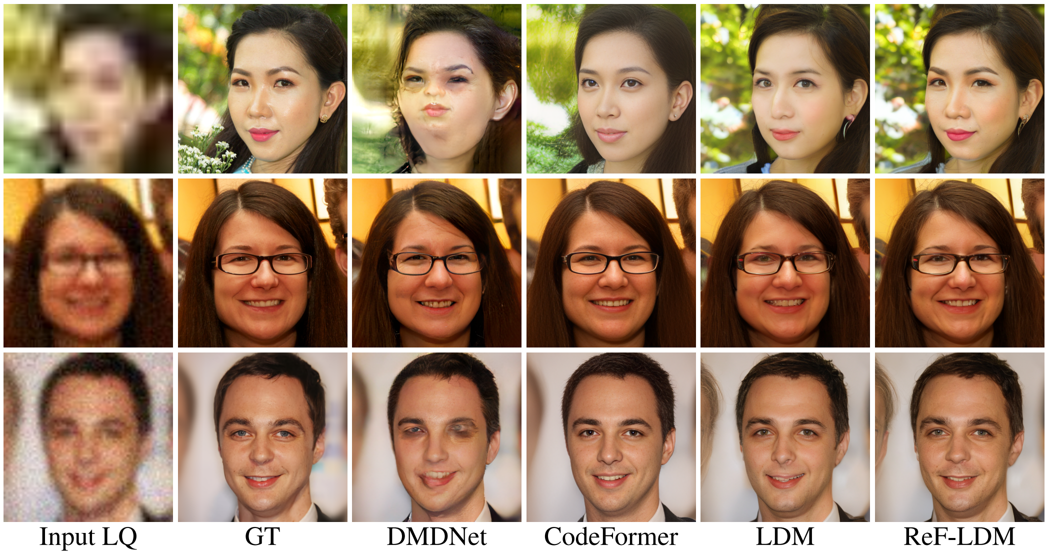


Figure 5: Qualitative comparison. From left to right: input LQ, ground truth, other methods, and our ReF-LDM. From top to bottom: FFHQ-Ref-Severe, FFHQ-Ref-Moderate, and CelebA-Test-Ref.

## 6 Limitations

When the face region is occluded by other objects, our model may generate artifacts. For certain face poses (e.g, side face), the reconstructed eyes may appear unnatural. These problems are also commonly observed in other methods and might be caused due to the lack of such training images. However, there are some examples showing that these problems can be alleviated if our model is provided reference images with similar face poses to the target image. Visual examples of these limitations are provided in Appendix F.

## 7 Conclusion

In summary, we propose ReF-LDM, which incorporates the CacheKV mechanism and the timestep-scaled identity loss, to effectively utilize multiple reference images for face restoration. Additionally, we construct the FFHQ-Ref dataset, which surpasses the existing dataset in both quantity and quality, to facilitate the research in reference-based face restoration. Evaluation results demonstrate that ReF-LDM achieves superior performance in face identity similarity over state-of-the-art methods.

## Acknowledgments

The authors wish to express their gratitude to Professor Wei-Chen Chiu for his valuable suggestion to exclude the reference images that are too similar to the target images when constructing the proposed FFHQ-Ref dataset.

## References

[1] M. Cao, X. Wang, Z. Qi, Y. Shan, X. Qie, and Y. Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22560–22570, October 2023.

[2] C. Chen, X. Li, L. Yang, X. Lin, L. Zhang, and K.-Y. K. Wong. Progressive semantic-aware style transformation for blind face restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11896–11905, 2021.

[3] J. Deng, J. Guo, X. Niannan, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019.

[4] P. Esser, R. Rombach, and B. Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021.

[5] Y. Gu, X. Wang, L. Xie, C. Dong, G. Li, Y. Shan, and M.-M. Cheng. Vqfr: Blind face restoration with vector-quantized dictionary and parallel decoder. In *European Conference on Computer Vision*, pages 126–143. Springer, 2022.

[6] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[7] J. Ho and T. Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[8] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[9] R. Huang, S. Zhang, T. Li, and R. He. Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. In *Proceedings of the IEEE international conference on computer vision*, pages 2439–2448, 2017.

[10] K. Karkkainen and J. Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1548–1558, 2021.

[11] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[12] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[13] Y.-F. Lau, T. Zhang, Z. Rao, and Q. Chen. Ented: Enhanced neural texture extraction and distribution for reference-based blind face restoration. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5162–5171, 2024.

[14] X. Li, M. Liu, Y. Ye, W. Zuo, L. Lin, and R. Yang. Learning warped guidance for blind face restoration. In *Proceedings of the European conference on computer vision (ECCV)*, pages 272–289, 2018.

[15] X. Li, W. Li, D. Ren, H. Zhang, M. Wang, and W. Zuo. Enhanced blind face restoration with multi-exemplar images and adaptive spatial feature fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2706–2715, 2020.

[16] X. Li, S. Zhang, S. Zhou, L. Zhang, and W. Zuo. Learning dual memory dictionaries for blind face restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5904–5917, 2022.

[17] X. Lin, J. He, Z. Chen, Z. Lyu, B. Fei, B. Dai, W. Ouyang, Y. Qiao, and C. Dong. Diffbir: Towards blind image restoration with generative diffusion prior. *arXiv preprint arXiv:2308.15070*, 2023.

[18] A. Mittal, R. Soundararajan, and A. C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012.

[19] Y. Nitzan, K. Aberman, Q. He, O. Liba, M. Yarom, Y. Gandelsman, I. Mosseri, Y. Pritch, and D. Cohen-Or. Mystyle: A personalized generative prior. *ACM Transactions on Graphics (TOG)*, 41(6):1–10, 2022.

[20] S. Pouyanfar, S. Sengupta, M. Mohammadi, E. Abraham, B. Bloomquist, L. Dauterman, A. Parikh, S. Lim, and E. Sommerlade. Frr-net: A real-time blind face restoration and relighting network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1240–1250, 2023.

[21] E. Richardson, Y. Alaluf, O. Patashnik, Y. Nitzan, Y. Azar, S. Shapiro, and D. Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2287–2296, 2021.

[22] 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*, pages 10684–10695, 2022.

[23] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4713–4726, 2022.

[24] J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[25] M. Suin, N. G. Nair, C. P. Lau, V. M. Patel, and R. Chellappa. Diffuse and restore: A region-adaptive diffusion model for identity-preserving blind face restoration. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6343–6352, 2024.

[26] Y.-J. Tsai, Y.-L. Liu, L. Qi, K. C. Chan, and M.-H. Yang. Dual associated encoder for face restoration. In *The Twelfth International Conference on Learning Representations*, 2024.

[27] J. Wang, Z. Yue, S. Zhou, K. C. Chan, and C. C. Loy. Exploiting diffusion prior for real-world image super-resolution. 2024.

[28] X. Wang, Y. Li, H. Zhang, and Y. Shan. Towards real-world blind face restoration with generative facial prior. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9168–9178, 2021.

[29] L. Zhang, A. Rao, and M. Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

[30] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

[31] Y. Zhao, T. Hou, Y.-C. Su, X. Jia, Y. Li, and M. Grundmann. Towards authentic face restoration with iterative diffusion models and beyond. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7312–7322, 2023.

[32] S. Zhou, K. Chan, C. Li, and C. C. Loy. Towards robust blind face restoration with codebook lookup transformer. *Advances in Neural Information Processing Systems*, 35:30599–30611, 2022.

## A Broader Impacts

The ReF-LDM has the capability to leverage personal appearances from reference images. This introduces a potential risk of misuse, where it could be employed for malicious face editing by using a low-quality image in conjunction with reference images from a different individual.

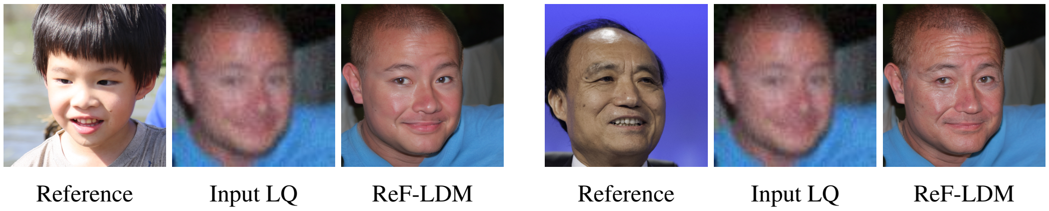


Figure 6: Examples of ReF-LDM using reference images from two different individuals.

## B Image quality issues in the previous dataset CelebRef-HQ

As described in Sec. 4.4, the previous dataset for the reference-based face restoration task, CelebRef-HQ [16], exhibits issues with image quality. We provide examples where the ground truth images in this dataset are corrupted by watermarks and mirror padding in Fig. 7.

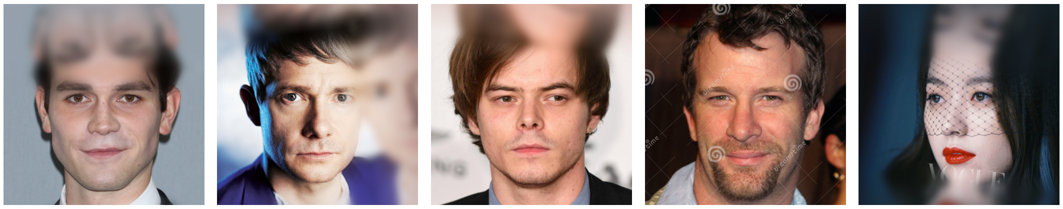


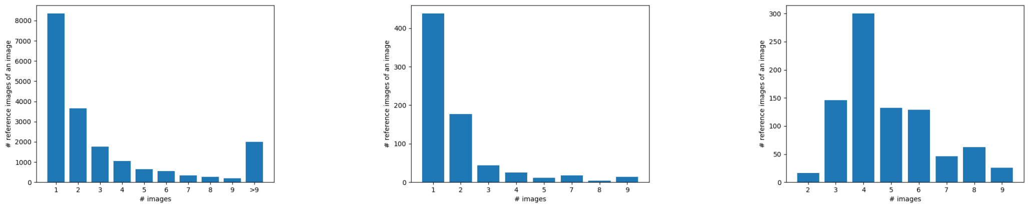
Figure 7: Example images with mirror padding and watermark artifacts in the CelebRef-HQ dataset.

## C Statistics of FFHQ-Ref dataset

We analyze the statistics of the proposed FFHQ-Ref dataset, introduced in Sec. 4.

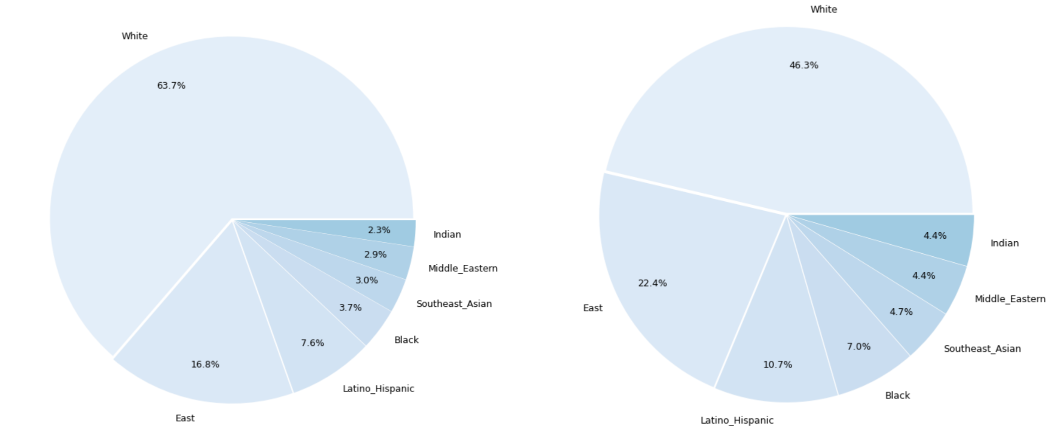
In Fig. 8, we plot the distribution of the number of available reference images.

Furthermore, we assess the race, age, and gender distributions of the dataset using labels predicted by FairFace [10]. As depicted in Fig. 9, the race distribution within FFHQ-Ref is imbalanced, with a predominance of the 'white' category. To mitigate this, we intentionally sampled a greater number of images from other races to construct a more balanced test set. Additionally, as illustrated in Fig. 10, FFHQ-Ref encompasses a broad age range, from infants (0-2 years) to the elderly (70+ years). However, the distribution is not uniform across ages and genders. For example, there is a notably higher proportion of young females (29.2% of '20-29 female').



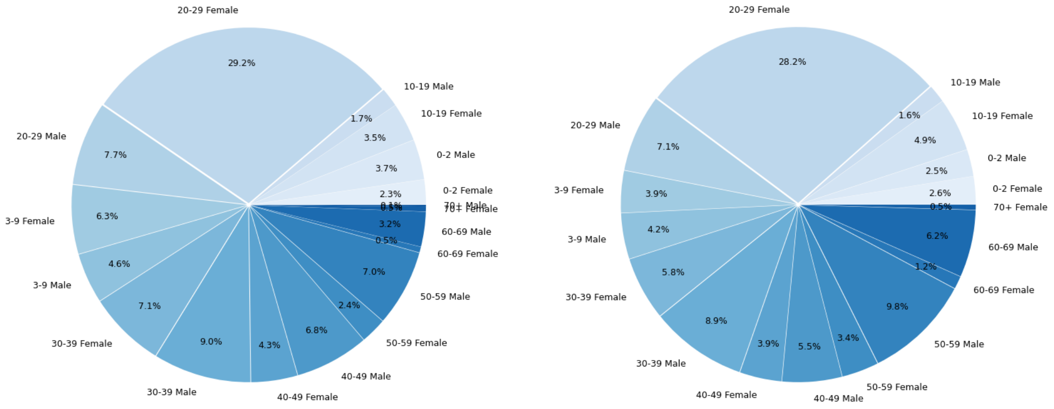
(a) Training set (b) Validation set (c) Test set

Figure 8: Distribution of the number of available reference images per image in the FFHQ-Ref dataset of train, validation, and test splits.



(a) FFHQ-Ref (b) FFHQ-Ref-Test

Figure 9: Race distribution within FFHQ-Ref dataset.



(a) FFHQ-Ref (b) FFHQ-Ref-Test

Figure 10: Age and gender distribution within FFHQ-Ref dataset.

## D Examples of differences in background regions

In Fig. 11, we provide some examples where our ReF-LDM are more different to the ground truth in background pixels compared to prior methods. In the first example, the ReF-LDM attempts to restore another face in the background. In the second example, our ReF-LDM restored the mirror padding in the CelebA-Test dataset as hairs.

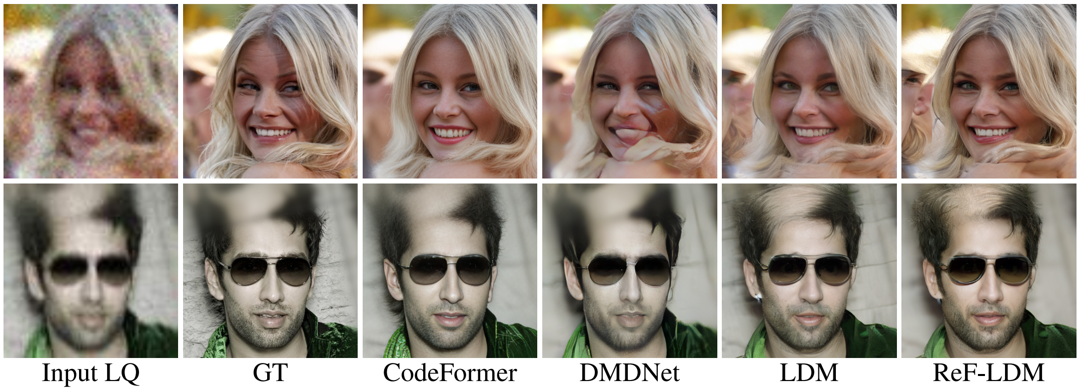
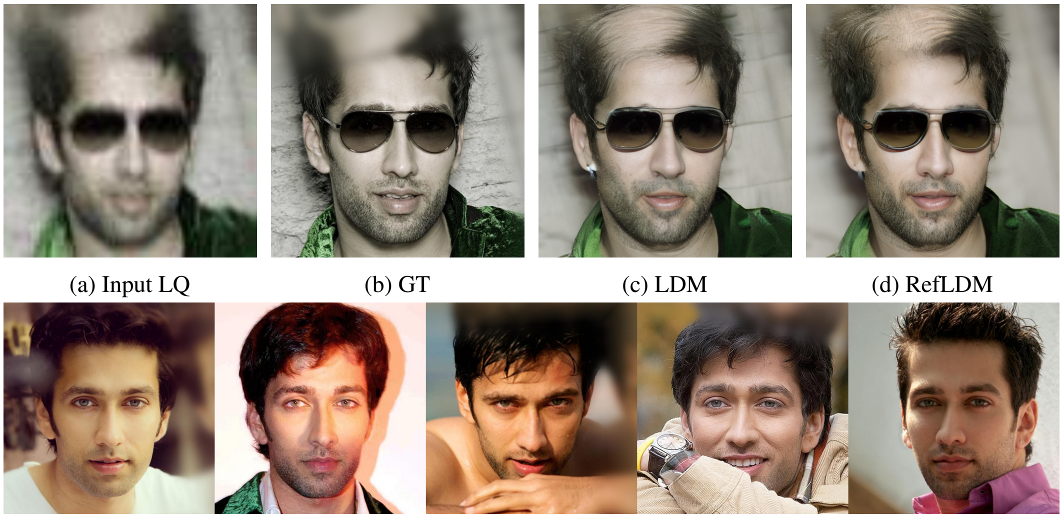


Figure 11: Examples where ReF-LDM generates background-region details that differ more from the ground truth.

## E Examples of illumination change

Fig. 12 shows an example where ReF-LDM exhibits warmer illumination compared to LDM. We conjecture that this may be due to the impact of the strong warm lighting in the input reference images. To address this issue, one could employ post-processing tricks, such as adjusting the means of the R, G, B channels to match those of the input LQ image. Another potential solution might be training ReF-LDM with data augmentation on the illuminations of input reference images, to encourage the model to disregard the illuminations of input references and maintain consistency with that of the input LQ image.



(a) Input LQ (b) GT (c) LDM (d) RefLDM

(e) Input references

Figure 12: An example of (d) ReF-LDM demonstrating an illumination change, likely influenced by the strong warm lighting of the (e) input reference images.

## F Examples of failure cases

Here we provide visual examples for the limitation described in Sec. 6. As shown Fig. 13, when the face region is occluded, our ReF-LDM and the prior models tend to generate from unnatural artifacts. For side face images, the ReF-LDM may not work well when the input reference images do not contain faces of similar pose, as shown in Fig. 14. However, Fig. 15 suggests that our ReF-LDM can effectively exploit the reference images of similar face poses to improve the results.

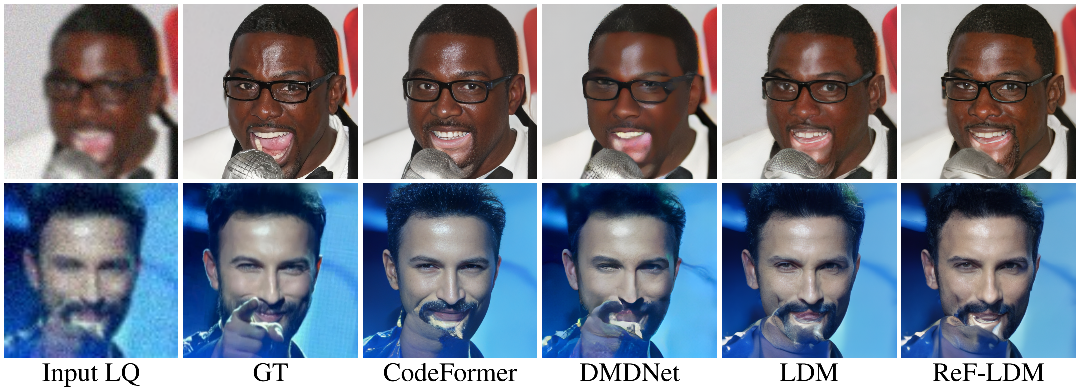


Figure 13: Some failure cases when the input LQ images are occluded.

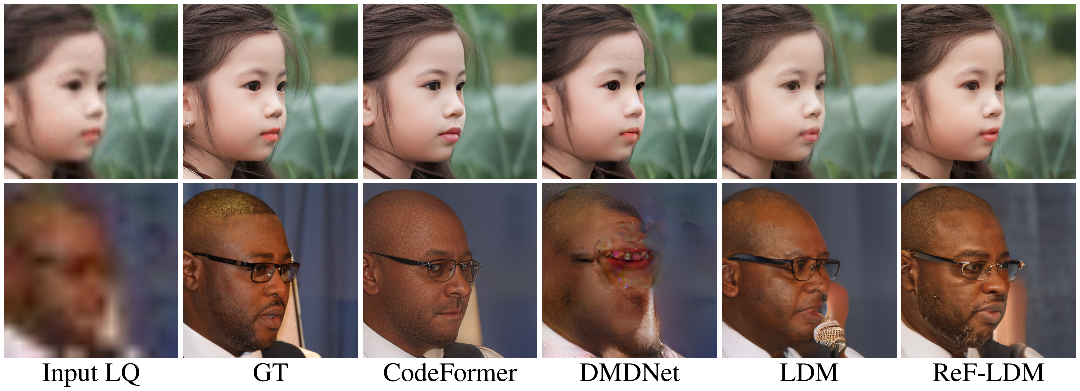


Figure 14: Some failure cases of side faces when side-face images are absent in the references.

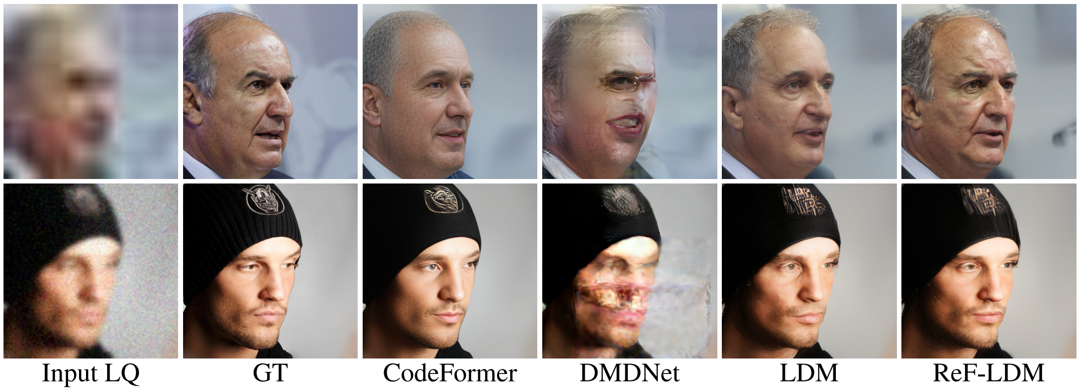


Figure 15: Some successful cases of side faces when side-face images are included in the references.

# G More implementation details

## G.1 Classifier-free guidance towards reference images

Classifier-free guidance [7] is a technique widely used in diffusion models for guiding the generated results towards a condition $c$ with a controllable scale factor $s$ at inference time:

$$\tilde{\epsilon}_\theta(\mathbf{z}_t, c) = \epsilon_\theta(\mathbf{z}_t, \varnothing) + s \cdot (\epsilon_\theta(\mathbf{z}_t, c) - \epsilon_\theta(\mathbf{z}_t, \varnothing)) \tag{5}$$

In our experiments, we use classifier-free guidance towards reference images with $s = 1.5$.

$$\tilde{\epsilon}_\theta(\mathbf{z}_t, \mathbf{z}_{\text{LQ}}, \{\mathbf{z}_{\text{ref}}\}) = \epsilon_\theta(\mathbf{z}_t, \mathbf{z}_{\text{LQ}}, \varnothing) + s \cdot (\epsilon_\theta(\mathbf{z}_t, \mathbf{z}_{\text{LQ}}, \{\mathbf{z}_{\text{ref}}\}) - \epsilon_\theta(\mathbf{z}_t, \mathbf{z}_{\text{LQ}}, \varnothing)) \tag{6}$$

During the training phase, we randomly drop the conditions by setting them to zero tensors with a probability of 0.1.

## G.2 Data augmentation for input reference images

During the training phase, we use a fixed number of five input reference images. When a target images with less than five reference images are sampled, we repeat the reference images to obtain five reference images. In addition, we apply image augmentation to the input reference images with the following operations: color jitter (brightness ± 0.2, contrast ± 0.2, saturation ± 0.2, hue ± 0.02), affine transform (rotation ± 2, translation ± 0.05, scale ± 0.05), perspective transform (scale ± 0.2, probability 0.5), and horizontal flip (probability 0.5). Lastly, we randomly shuffle the order of available reference images for a target image, so that a different combination of reference images can be sampled at each training iteration. In Fig. 16, we provide an example where a set of two reference images is augmented to a set of five reference images.

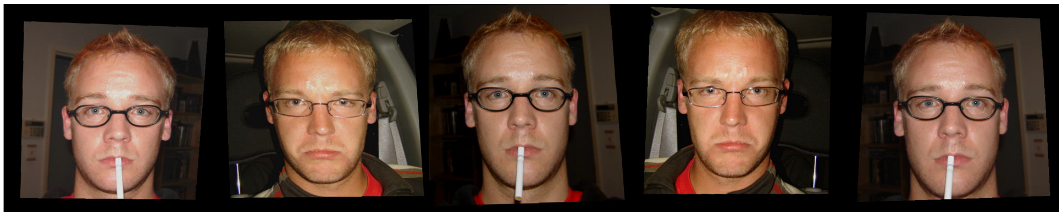


Figure 16: Data augmentation for reference images.

## G.3 Training details

We trained the VQGAN for 200,000 iterations with batch size 32 on four A6000 GPUs for 7 days. We trained the LDM with only LQ condition for 500,000 iterations with batch size 40 on four A6000 GPUs for 7 days. We finetuned the ReF-LDM for 150,000 iterations with batch size 8 on four 3090 GPUs for 6 days. For training losses, the LDM is trained using only the typical LDM loss $\mathcal{L}_{\text{LDM}}$, while the ReF-LDM is trained with both $\mathcal{L}_{\text{LDM}}$ and the proposed $\mathcal{L}_{\text{time ID}}$.

### G.4 Hyperparameters of networks

For the frozen autoencoder, we use a VQGAN as in the LDM [22] with the following settings:

- input image: 512x512x3
- latent representation: 64x64x8
- code booksize: 8192
- network hyperparameters: base channel as 128, multiplier for each scale as [1, 1, 2, 4] with 2 residual blocks.

For the denoising U-net, we use the following settings:

- input latent: 64x64x16
- output latent: 64x64x8
- attention layer at resolutions: 32x32, 16x16, and 8x8
- network hyperparameters: base channel as 160, multiplier for each scale as [1, 2, 2, 4] with 2 residual blocks.