Personalized Restoration via Dual-Pivot Tuning

Pradyumna Chari, Sizhuo Ma, Daniil Ostashev,
Achuta Kadambi, Gurunandan Krishnan, Jian Wang, Kfir Aberman

***Abstract*—Generative diffusion models can serve as priors, ensuring that image restoration solutions adhere to natural image manifolds. For facial images, however, personalized priors are essential to accurately reconstruct individual-specific facial features. We propose *Dual-Pivot Tuning* — a simple yet effective two-stage approach to personalize blind restoration systems while preserving general prior integrity. Our key observation is that for efficient personalization, the diffusion model should be tuned around a fixed textual pivot in the first step, while in the second step a guiding network should be tuned in a generic (non-personalized) manner, using the personalized diffusion model as a fixed "pivot". This approach ensures that personalization does not interfere with the restoration process, producing results with a natural appearance that show high fidelity to both identity and degraded image attributes. We conducted extensive experiments with images of widely recognized individuals, evaluating our approach both qualitatively and quantitatively against relevant baselines. Notably, our personalized prior not only achieves superior identity fidelity, but also outperforms state-of-the-art generic priors in terms of overall image quality. Project webpage is https://personalized-restoration.github.io/ and code is available at https://github.com/personalized-restoration/personalized-restoration.**

***Index Terms*—Text-guided personalization, face restoration, generative priors**

I. Introduction

IMAGE restoration is an extensively researched problem, characterized by its inherent ill-posed nature. The goal of the restoration task is to find visually plausible and natural images that maintain perceptual fidelity to the degraded input image [3]. In blind restoration scenarios where no prior information about the subject or the degradation is available, a prior describing the manifold of natural images is needed. However, for face image restoration, having an identity prior is necessary to ensure that the output image remains within a manifold that accurately represents the distinctive facial features of the individual in the degraded image. This sets the basis for prior work in reference-based face image restoration [4]–[6].

On a parallel track, text-to-image diffusion models [7], [8] have revolutionized image synthesis, enabling the generation of images from textual descriptions. These models act as versatile generative priors for various downstream tasks, and were recently explored in the context of blind image restoration, enhancing the naturalness of restored images [1], [9]. Moreover, the ability to personalize these diffusion models with just a few reference images opened up new avenues for tailored content creation [2], [10]. Despite these advancements, effectively integrating personalization into diffusion-based blind restoration systems remains an open problem. To highlight the challenges associated with this problem, Fig. 2 demonstrates how naive personalization approaches fall short. For instance, a personalized text-to-image model that simply replaces the general prior in a blind restoration system (from [1], for example) can be disregarded by the unconditional system which returns an output that resembles the result of the blindly restored one (Fig. 2 (c)). Alternatively, naive personalization of the encoder, which is equivalent to how the system was trained, disrupts the natural image prior, leading to a lack of detail (Fig. 2 (d)).

In this paper, we introduce a simple and efficient technique for personalized image restoration. Given a small set ($\sim$10) of high-quality images of a person, we aim to restore their degraded images while ensuring that the restored image shows: (1) strong identity preservation, (2) high fidelity to the degraded image input, and (3) natural visual appearance. Our approach, termed *dual-pivot tuning*, personalizes a blind image restoration system comprising two core components: a diffusion-based generative prior and a guiding image encoder. This methodology maintains both the integrity of the underlying prior and preserves fidelity to the attributes present in the guiding images.

To achieve that, we propose a two-step solution. In the first step, we personalize the diffusion-based generative prior, while using a *textual pivot*—a fixed, unique, token within a text-prompt (for example, "a photo of a [v] man") that is held constant during the fine-tuning process [2]. However, in our case, the fine-tuning is performed with the guiding image encoder such that the personalized prior learns to respect the attributes of the guiding image. In the second step, we fine-tune the guiding encoder to align with the personalized, "shifted", prior of the diffusion model, thereby retargeting its functionality. During this phase, we aim to maintain the identity-agnostic nature of the encoder, allowing its applicability across various personalized models. Therefore, we fine-tune it with general face images (not of the specific individual). We refer to the fixed personalized network as a pivot, around which we adjust the weights of the guiding encoder.

We find both of these operations, in this sequence, to be essential. The textual pivoting enables identity injection

P. Chari and A. Kadambi are with University of California, Los Angeles, CA 90095, USA. (e-mails: pradyumnac@g.ucla.edu, achuta@ee.ucla.edu)

S. Ma, D. Ostashev, G. Krishnan, J. Wang, and K. Aberman are with Snap Inc., Santa Monica, CA 90405, USA. (e-mails: sma@snap.com, dostashev@snap.com, guru@gurukrishnan.com, jwang4@snap.com, kaberman@snap.com)

This work was partially done while P. Chari was an intern at Snap Inc.

J. Wang is the corresponding author.

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the author. The material includes additional experimental resuts and analysis. Contact jwang4@snap.com for further questions about this work.

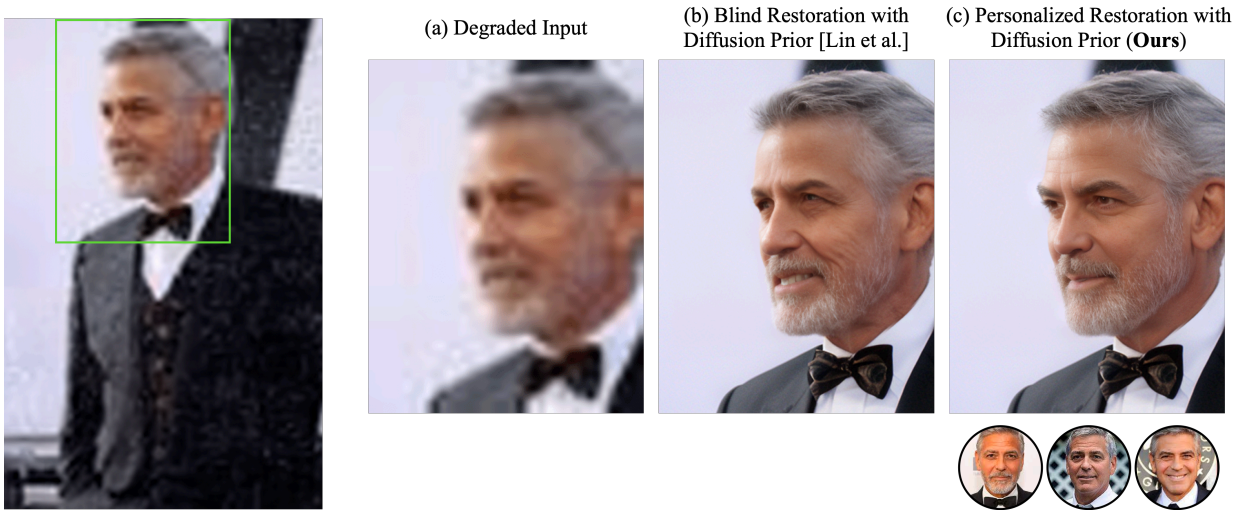


Fig. 1: Given a degraded image of an individual's face (a), diffusion-based blind restoration approaches [1] may not retain the individual's identity (b). However, with a few reference images (bottom right), our *dual-pivot tuning* technique (c) can reconstruct the face while maintaining high identity fidelity to the individual without perceivable loss in fidelity.

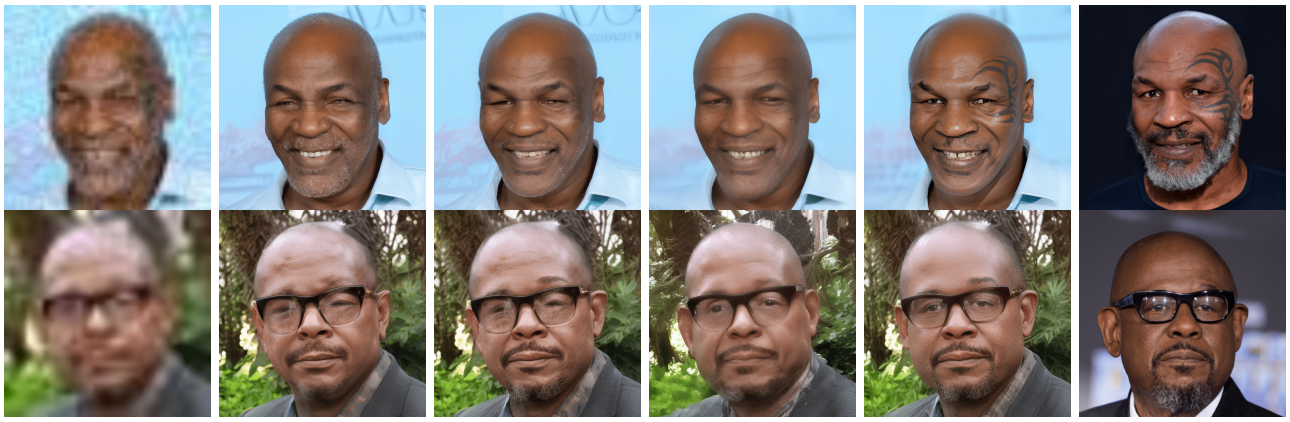


Fig. 2: **Restoration Baselines.** Given a real, degraded input image (a) features a person whose identity is referenced in another image (f), a diffusion-based blind restoration method [1] with a general (non-personalized) prior, fails to preserve identity (b). When fine-tuning independently the text-to-image prior with a text pivot only [2], the system ignores the personalized generator when reintegrated into the system (c). Naively fine-tuning the system (fine-tuning the encoder while holding the generative model fixed) leads to lack of detail (tattoo in the first row, beard in the second row), along with absence of generalization across identities (d). In contrast, our method (e) effectively incorporates the individual's identity into the restoration process while retaining quality comparable to the base model (b).

without losing the general face prior of the base model in the restoration system, and fine-tuning around the model-based pivot leads to better utilization of the guiding encoder and higher fidelity to the input image features. Furthermore, we leverage the diffusion process's characteristics, noting that identity formation occurs later in the process, to reduce the expansive fine-tuning time by $\sim \times 2$, which can be a significant reduction in cost in a large user base.

We compare our method against reference-based baselines and evaluate them using publicly available images of well-known figures, leveraging our pre-existing knowledge of their facial features. Our experiments show that our method reconstruct key facial features of the subject in the reference images while maintaining high fidelity to the original degraded image, outperforming other methods both quantitatively and qualitatively. Moreover, our user study confirms that the personalization contributes significantly not only to identity preservation but also to the overall improvement of perceptual quality. Fig. 1 shows one example that demonstrates how our method can surpass an existing diffusion-based blind face restoration method [1] in terms of identity preservation while maintaining the quality of face restoration. Additionally, the text-guided, multimodal nature of our approach enables secondary applications such as text-guided editing.

II. Related Work

A. Blind Face Image Restoration

Distinct from general scene image restoration, face image restoration typically leverages facial priors to achieve superior results. Depending on how these facial priors are utilized, previous methods can be broadly categorized into three main classes. (1) Geometric prior: These approaches incorporate geometric cues, such as facial landmarks [11]–[13], parsing maps [14]–[16], and component heatmaps [17], into the network design. (2) Dictionary prior: A dictionary is first learned from a collection of face images, either in the image space [18] or a feature space [19]–[22]. Subsequently, a degraded image is restored using high-quality words from this dictionary. (3) Generative prior: This category encompasses techniques like GAN (Generative Adversarial Network) priors [23]–[26] and diffusion priors [1], [27]–[29]. Among these methods, approaches falling under categories (2) and (3) have demonstrated the most promising results. Notable algorithms include GFP-GAN [25], CodeFormer [22], and DiffBIR [1].

However, blind face image restoration faces a significant challenge known as the quality-fidelity tradeoff [1], [22]. This arises from the inherent limitations in the information available in the original image. Striking the right balance between generating high-quality results while staying faithful to the original image can be a delicate task. Generating results with too little modification may not yield an improvement in quality, while excessive generation can lead to a departure from the identity of the original image. Our proposed method differs from these: using personalized diffusion-based methods, we are able to improve the tradeoff by achieving restoration quality while retaining fidelity with respect to identity.

B. Reference-Based Face Image Restoration

A high-quality reference image of the same person can greatly benefit face image restoration and help avoid the need of the tradeoff. Depending on the number of references used, such methods can be divided into two categories. (1) Single-reference methods, including GFRNet [30], GWAINet [31], and ReFine [32]. (2) Multi-reference methods, including ASFFNet [4], Wang *et al*. [33], DMDNet [5] ($\leq$ 10 reference images), and MyStyle [6] ($\sim$ 100 images). It is evident that multi-reference approaches yield superior results as they leverage more information. Notably, ASFFNet [4] selects an optimal guidance, Wang *et al*. [33] employs pixel-wise weights for multiple references, and DMDNet [5] constructs a dictionary from multiple references; [6], [34], on the other hand, fine-tunes StyleGAN based on personal images.

Our proposed method also utilizes multiple reference images to aid personalized restoration. However, there are several distinctions. We use a diffusion-based personalized generative prior, while [4], [30], [31], [33] use feedforward architectures. MyStyle [6] uses a GAN-based generative prior, but it requires around 100 images for effective personalization and strict spatial alignment of the face landmarks within each image. In contrast, our method only needs 10 reference images and has no restriction on the alignment. This leads to higher quality restoration with less restrictions for our proposed method.

C. Personalized Diffusion Models

Diffusion models [35]–[37] have notably excelled in the area of generating images from text (T2I) [7], [8], [38] and many other visual computing tasks [39]. Recent advancements in this field involve customization of these established models through fine-tuning, aiming to enhance features like controllability, customization, or to cater to specific applications. One approach to customization involves modifying the T2I model itself [2], [40]–[42] or the text embedding process [10], [43], [44], using selected images. This allows to personalize the generation of images based on a particular subject or style, driven by textual input. More recently, neurons in a model specific to a concept can be identified and manipulated specifically to enable sparse personalization [45], methods for fast personalization of the model have been proposed [46]–[48], and approaches for multi-subject personalization [49], [50]. Alternatively, other methods involve adapting the T2I model to introduce new conditioning factors. These changes are either for purposes of image modification [51]–[53] or for generating more controlled images [54], [55].

In this work, we tackle the task of contextual customization, where the goal is to fine-tune a prior within the context of a system, preserving the distinct roles of each component within the system, while customizing the image prior to a specific subject.

III. Preliminaries

A. Personalizing Text-Guided Diffusion Models

Our method relies on multimodal text-guided image generation models. Given a text prompt $\mathbf{p}$, text-guided denoising diffusion models learn to sequentially denoise samples of random noise $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ into samples of images $\mathbf{x}$. During training, a neural network $\epsilon_\theta(\mathbf{x}_t, \mathbf{p})$ is trained to predict ϵ from a noisy version of the image $\mathbf{x}_t = \alpha_t \mathbf{x} + \sigma_t \epsilon$, where α_t and σ_t are noise scheduling parameters, and t refers to the time step in the diffusion process.

A common way to sample the image at inference time is classifier-free guidance [56], that sums up a conditional instance of the model together with an unconditional one:

$$\mathbf{G}_\theta(\mathbf{x}_t, \mathbf{p}) = (1 + w)\epsilon_\theta(\mathbf{x}_t, \mathbf{p}) - w\epsilon_\theta(\mathbf{x}_t, \varnothing), \tag{1}$$

where w is a guidance scale, $\varnothing$ represents a null-text prompt. These conditional and unconditional branches aim to strike a balance between prompt fidelity and diversity within the generated images.

Recent work by [2] introduced the ability to personalize text-guided diffusion models using a few reference images of a subject. This personalization process involves fine-tuning the model around a text anchor $\mathbf{p}$ (typically structured as a "rare token"+"class noun", such as *'a [v] dog'*), enabling subject identity embedding within the diffusion prior that can be activated when $\mathbf{p}$ appears in the conditioning during inference. In practice, the weights, θ, are optimized by minimizing

$$\mathcal{L}_{\text{DB}} = \mathbb{E}_{\mathbf{x},\mathbf{p},\epsilon,\epsilon',t}\left[||\mathbf{G}_\theta(\mathbf{x}_t, \mathbf{p}) - \epsilon||_2^2 + ||\mathbf{G}_\theta(\mathbf{x}_t^{\text{pr}}, \mathbf{p}^{\text{pr}}) - \epsilon'||_2^2\right], \tag{2}$$

where $\mathbf{x}_t^{\text{pr}}$ is drawn from a separate prior preservation dataset, and $\mathbf{p}^{\text{pr}}$ is the prior prompt ("class noun", such as *'a dog'*).

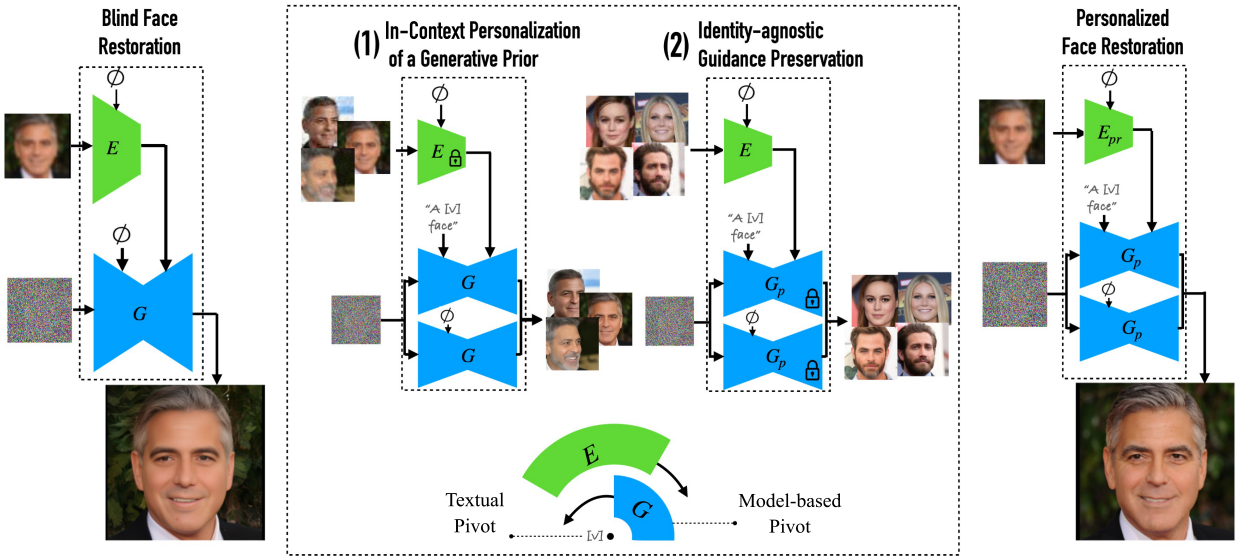


Fig. 3: **High-level overview.** Our approach aims to personalize a blind face restoration system (left) in two steps: (1) text-anchored personalization of the generative prior $\mathbf{G}$ within the context of the system, and (2) retargeting the encoder $\mathbf{E}$ in the presence of the personalized prior in $\mathbf{G}_p$, in an identity-agnostic manner. Then, at inference time (right), our system can generate output images with high fidelity to the individual appearing in the reference images, while retaining the perceptual fidelity to the degraded input.

(a) Input (b) Seed 1 (c) Seed 2 (d) Seed 3

Fig. 4: **Identity variation.** Given a fixed degraded input image, it can be seen that different seeds (i.e. different input noises) applied to [1] result in images with different identities.

This structure allows the optimization to revolve around a text pivot $\mathbf{p}$, without ruining the general prior of the model [57].

We define the personalization operator of a generative model by

$$\mathbf{M}_{\mathcal{W}} = \mathcal{P}\{\mathbf{M}, \mathcal{W}\}, \tag{3}$$

where $\mathcal{P}\{\cdot\}$ is the operator, and $\mathbf{M}_{\mathcal{W}}$ is the personalized version of a model $\mathbf{M}$ which is fine-tuned with $\mathcal{W}$ as a pivot. That is, the first argument of the operator is finetuned while keeping the second argument fixed. In our case, this pivot can be either a text prompt or a network.

IV. Method

In this section, we outline our method for personalizing guided diffusion models. We begin by describing the design and operation of the diffusion-guided blind image restoration system. Next, we present our dual-pivot tuning technique, which involves two key steps: first, employing text-based fine-tuning to embed identity-specific information within diffusion priors, and second, addressing the necessity of model-centric pivoting to adjust the guiding image encoder with the integrated personalized priors. Finally, we show how the inherent characteristics of our method can be utilized to accelerate the per-identity fine-tuning process and to significantly reduce the cost of the process.

A. Diffusion-Guided Image Restoration

Recent advances in blind face image restoration [1], [9] have attempted to utilize a pre-trained diffusion model (such as Stable Diffusion [7]) as generative prior to guide the restoration process. Typically, these pipelines contain a diffusion-guided image restoration step that consist of two components: (1) A general image prior in the form of a text-to-image diffusion model $\mathbf{G}(\cdot)$, that encodes the manifold of natural images and (2) an encoder $\mathbf{E}(\cdot)$, that captures context information from the degraded image and guides the generation process such that it maintains high fidelity to the visual attributes of the degraded image. Guiding the diffusion process with spatial features extracted from an image encoder $\mathbf{E}$ is a common way to control the diffusion processes [54]. It should be noted that in the case of blind restoration, the general prior $\mathbf{G}$ plays a key role in dictating the identity of the person. As demonstrated in Fig. 4, for a given degraded input image (namely, fixed set of guiding features that are extracted from $\mathbf{E}$), different seeds in the input of $\mathbf{G}(\cdot)$, lead to different identities in the output.

Our goal in this work is to personalize the general prior $\mathbf{G}$ such that the restored face maintains high fidelity to the identity of the individual without degrading the integrity of the general image prior and the distinct roles of each component in the system.

We consider a general setting, where the encoder $\mathbf{E}(\cdot)$ takes a low-quality image I_{LQ} as an input and provides guidance for the diffusion model. Note that I_{LQ} could be a degraded image [9], or the output of a preliminary restoration model,

like in [1]. In the context of the blind image restoration setting, we will use the following notation:

$$\mathbf{B} = \{\mathbf{G}(\epsilon, \varnothing), \mathbf{E}(I_{\text{LQ}})\}, \tag{4}$$

where $\mathbf{B}$ represents the blind restoration system with the two aforementioned components. In this blind restoration setting, the text conditioning is null, denoted by $\varnothing$ (as in [1]).

B. Dual-Pivot Tuning

A high-level summary of our two-step approach is illustrated in Fig. 3.

a) Step 1: In-Context Textual Pivoting: Given a pivot text prompt p of the form "A [v] face", personalization of a generative prior in a restoration system can be performed by fine-tuning $\mathbf{G}$ with p independently of the blind restoration system and plug it back into the framework, namely,

$$\mathbf{B}_p = \{\mathbf{G}_p, \mathbf{E}\}, \tag{5}$$

where $\mathbf{G}_p = \mathcal{P}\{\mathbf{G}, p\}$. However, as demonstrated in Fig. 2 (c), if we do so, the personalized prior is being completely ignored by the system. Alternatively, we can personalize $\mathbf{G}$ within the context of $\mathbf{B}$, while pivoting around the prompt p, such that it leverages the conditioning cues from $\mathbf{E}$ during fine-tuning, namely,

$$\mathbf{B}_p = \mathcal{P}\{\mathbf{B}_{\mathbf{G}}, p\}, \tag{6}$$

where $\mathbf{B}_{\mathbf{G}}$ denotes that we modify the weights of $\mathbf{G}$ within the context of $\mathbf{B}$ (i.e. modify the generator within the blind restoration system). We refer to this as in-context pivoting. While restoration is better comparing to the decoupled (out-of-context) fine-tuning, the restored images are unable to completely imbibe identity information and fidelity cues, when compared to our final result, as demonstrated in Fig. 9.

The occurrence of this issue can be attributed to the initial training of $\mathbf{E}$, where it was coupled with the unconditional branch of $\mathbf{G}$. As a result, $\mathbf{E}$ cannot exert influence on the conditional branch of $\mathbf{G}$, which is responsible for contributing personalized features around the anchor prompt p. We therefore find that for effective personalization of face image restoration, text-based pivoting alone is not enough, and that $\mathbf{E}$ must be fine-tuned in conjunction with $\mathbf{G}_p$.

b) Step 2: Model-based Pivoting: We introduce a second fine-tuning step for $\mathbf{E}$ within the context of $\mathbf{G}_p$. Essentially, our objective is to fine-tune $\mathbf{E}$ in a way that it relinquishes identity cues to the personalized prior in $\mathbf{G}_p$ while focusing on other detail cues. In this scenario, our pivot is not a text but a network ($\mathbf{G}_p$), which we keep fixed during the optimization process to align the encoder with the personalized prior.

To preserve the identity-agnostic role of the guiding encoder, we intentionally aim to avoid personalizing it. One approach to achieve this is by updating $\mathbf{E}$ across different individuals, allowing it to adapt to the conditional branch of $\mathbf{G}_p$ while remaining agnostic to fine-grained identity features, which should be determined by $\mathbf{G}_p$, as demonstrated in Fig. 4. This identity-agnostic retargeting concept bears similarities to the Prior Preservation loss [2] in the context of pivotal-tuning around a model. We denote this process as identity-agnostic guidance preservation.

Starting with $\mathbf{G}_p$, we perform fine-tuning on $\mathbf{E}$ as follows:

$$\mathbf{B}_p = \mathcal{P}\{\mathbf{B}_{\mathbf{E}} = \{\mathbf{G}_p, \mathbf{E}\}, \mathbf{G}_{\mathbf{p}}\}, \tag{7}$$

where $\mathbf{B}_{\mathbf{E}}$ represents the restoration system (with the embedded personalized prior), and we are now optimizing the weights of $\mathbf{E}$. The encoder obtained as a result of this step is denoted as $\mathbf{E}_{\text{pr}}$.

Combining these two steps is crucial. The textual pivot allows for identity injection without altering the base model's prior in the context of the restoration system, while fine-tuning around the network pivot results in better use of the guiding encoder and greater fidelity to the input image features.

Note that the unique identifier [v] is specifically associated with the generative prior $\mathbf{G}$ and forms an essential part of the in-context personalization of $\mathbf{G}$. Therefore, during ID-agnostic training, the identifier [v] that is associated with the identity of the input image is passed to $\mathbf{G}$. This is distinct from the identity-agnostic guidance preservation, which operates on the encoder $\mathbf{E}$ and is conditioned with a null text prompt. Using an identity-specific identifier for $\mathbf{G}$ is crucial because the encoder $\mathbf{E}$ is required to remain identity-agnostic. This conditioning structure is fundamental to the effectiveness of our approach in preserving both identity-specific features and generalized restoration capabilities.

C. Speeding-Up Dual-Pivot Tuning

In general, personalization of generative models is a time-consuming process. We next suggest steps to significantly reduce (by about 2x) fine-tuning time in each of the two phases.

a) Speeding up textual pivoting.: The presence of both conditional and unconditional branches within our inference pipeline offers an additional unique opportunity. We noticed that the initial restoration steps do not rely as much on identity, as these steps mainly focus on coarse geometry and semantic detail [44]. Therefore, during inference, we observe that high-quality, identity-preserving restoration can be achieved even when the initial denoising steps are only unconditional, using a guidance scale of 0, followed by text-guided denoising for later steps with a non-zero guidance scale. Formally,

$$\begin{aligned} \epsilon_{t+1} =& \mathbf{B}_p(\mathbf{x}_t, \varnothing) \mid 0 \leq t < \gamma, \\ \epsilon_{t+1} =& w \cdot \mathbf{B}_p(\mathbf{x}_t, p) + (1-w)\mathbf{B}_p(\mathbf{x}_t, \varnothing) \mid \gamma \leq t < 1.0, \end{aligned} \tag{8}$$

where γ denotes the fraction of denoising steps for which unconditional inference is carried out. Note that in the above expression, we denote the initial noise to be at $t = 0$ and the restored image to be at $t = 1$.

We demonstrate this in Fig. 5, where we restore the degraded image (a) unconditionally for the first $\gamma = 50\%$ of the steps and conditionally for the remaining steps (c). Despite this, we find the restored image being faithful to the reference image (d), as well as to the restored image obtained through conditional restoration for all the steps (b). $\gamma = 50\%$ is identified experimentally. This observation enables us to reduce personalization time for the textual pivoting by about

(a) (b) (c) (d)

Fig. 5: **Unconditional Sampling.** During inference, identity-preserved restoration of a degraded image (a) is possible even with unconditional denoising for initial steps, followed by conditional denoising for remaining steps (c). We find this to be consistent with the reference image (d), in terms of identity and faithfulness to original image, as well as with the result through conditional restoration for all steps (b).

(a) Input (b) $\mathbf{E}$, same id. (c) $\mathbf{E}$, diff. id. (d) ID Reference

Fig. 6: **Finetuning $\mathbf{E}$ on different identities.** We show that for the model-based pivot tuning $\mathbf{E}$ can be finetuned on the same identity (b), as well as on different identities as in the inset (c), while providing similar plausible restorations with respect to the identity in the reference image (d).

2x since the denoiser no longer needs to be personalized for higher noise levels (since the unconditional model suffices for those), and can, therefore, be trained with a focus on relatively lower noise levels.

b) Speeding up model-based pivoting.: We find that in-context textual pivoting (Section IV-B) is effective towards speeding up model-based pivoting. With the personalization of the generative prior being in-context in our proposed pipeline, model-based pivoting is feasible with half the number of finetuning steps, when compared with an out-of-context textual pivoting pipeline (Fig. 9). A more detailed runtime analysis for the speedup may be found in Table II.

Moreover, we have observed that the personalized prior in $\mathbf{G}_p$ is sufficiently strong, such that even when performing the model-based pivoting of $\mathbf{E}$ on a single individual, $\mathbf{E}_{\text{pr}}$ captures general cues and can be shared across various identities, as demonstrated in Fig. 6. This is critical from a training time perspective: the fine-tuned encoder, once performed on one identity, can be shared across models for different identities. Therefore, the only personalization step needed for each new identity is the textual pivoting.

While it might initially appear that ID-agnostic guidance preservation is weaker in the context of Figure 6 since training is performed on only one identity, this is not the case. As demonstrated in Figure 6(b) and (c), even through this seemingly constrained training process, the encoder successfully maintains ID-agnostic properties. It can be used across multiple identities while retaining strong image restoration capabilities. This indicates that our structured, two-step pipeline allows training of an ID-agnostic guidance encoder without requiring images from diverse identities. This substantially increases the practical utility of our method, for example in scenarios where privacy concerns might limit the availability of training images across diverse identities.

These observations allow us to efficiently save a significant amount of computing time, ultimately resulting in substantial cost reductions, especially when dealing with a large-scale user base. The final method we run for our experiments includes these critical speedup steps, including the single identity model-based pivoting of $\mathbf{E}$.

V. Experiments

In this section, we highlight the superiority of the proposed method to prior reference-based and blind image restoration methods.

A. Training Process and Datasets

We use images from CelebRef-HQ dataset [5] to personalize the model; more specifically, this dataset has multiple 512×512 images of the same person. Our method is trained on synthetic data that covers a wide range of degradation similar to the real world. We use DiffBIR [1] as our base blind image restoration framework. In this setting, the generative prior $\mathbf{G}$ is Stable Diffusion v2.1 [7], while the encoder $\mathbf{E}$ is initialized as the image-conditional encoder from DiffBIR. Specific details on training, architecture and so on may be found in Appendix C.

We follow the same second-order degradation strategy as [1]. At each stage we first convolve the image with a blur kernel k_σ, and downsample with a scale factor r. Following that, additive noise n_δ is added, and finally the image is JPEG-compressed with quality q. Formally, a single stage can be described as

$$\mathbf{x}_d = [(\mathbf{x} \circledast k_\sigma) \downarrow_r + n_\delta]_{\texttt{JPEG}_q}, \tag{9}$$

where $\circledast$ is the convolution operator. The final image is obtained after applying Eq. (9) twice. We refer to [1] for more details on the degradation process.

a) Test data.: We use two sources of data: (1) CelebRef-HQ test set, (2) Google searched images including low-quality and high-quality images from the same person. Real-world degradations are an arbitrary unknown combination of image compression, low resolution, image blur and noise.

B. Comparisons with other strategies

We show the comparisons of our personalization strategy to the alternative strategies in Fig. 2. As can be seen, the proposed strategy maximizes identity preservation through the restoration process, while comparison strategies, namely personalizing when pivoting only on the text condition (c) or personalizing when pivoting only on the generative model both are unable to effectively incorporate identity while retaining image fidelity.

(a) DEGRADED IMAGE (b) ASFFNET [60] (c) DMDNET [5] (d) DIFFBIR [1] (e) OURS (f) GROUND TRUTH

Fig. 7: **Results on synthetically degraded images.** A considerable gap in identity preservation can be observed between our proposed method and alternative baselines. Identity preservation can be observed in terms of overall geometric features, as well as attributes: eye shape (row 1), nose and nostril shape (row 2), eye color (row 3). Please also explore the webpage included with the supplementary material for an interactive side-by-side comparison, a better visualization of the differences.

TABLE I: **Fidelity and identity metrics.** The proposed method generates images with high fidelity, while showing superior identity retention, when personalized on 10 images. We compare along PSNR, SSIM, LPIPS and FID as fidelity metrics, while ArcFace similarity serves as an identity metric. We highlight the **best**, second-best and *third-best* performing methods.

Method	PSNR (dB)	SSIM	LPIPS	FID	ArcFace (Identity)
GFPGAN [25]	23.67	0.5783	0.1041	*59.76*	*0.75*
CodeFormer [22]	23.98	*0.5726*	**0.0920**	58.50	0.76
DPS [58]	23.16	0.5373	*0.2049*	107.40	0.42
DiffFace [59]	23.05	0.5288	0.3743	201.31	0.54
MyStyle [6]	17.24	0.5265	0.2615	103.23	0.68
DR2 [28]	21.56	0.5433	0.2132	113.60	0.40
ASFFNet [60]	11.21	0.4042	0.3070	186.54	0.55
DMDNet [5]	11.33	0.4010	0.3209	176.19	0.61
DiffBIR [1]	**24.27**	**0.5797**	*0.1031*	65.44	0.76
Ours	*23.72*	0.5532	0.0949	**57.88**	**0.88**

C. Comparisons with SOTA methods

We perform qualitative and quantitative comparisons. For reference-based methods, we choose ASFFNet [4], DMDNet [5], MyStyle [6], as well as DiffFace [59], a face swapping-based method; all methods use 10 reference images for guidance. For blind face image restoration, we chose DiffBIR [1], DR2 [28], GFP-GAN [25], CodeFormer [22] and diffusion posterior sampling (DPS) [58]. **Please find qualitative comparison with MyStyle, DR2, GFP-GAN, CodeFormer in the supplement, and comparisons with DiffFace [59] and DPS [58] in the Supplement.**

a) Comparison on synthetic degradations: Fig. 7 (and additional results in the supplement) show qualitative comparisons on the CelebRef-HQ test set subject to the synthetic degradations described above. It is noteworthy that ASFFNet and DMDNet are trained on perfectly aligned face images. During test time, highly accurate face alignment is required to eliminate domain gap. As a result, the results of both methods are underwhelming. On the other hand, diffusion-based methods (DiffBIR and ours) possess a generative prior trained on huge amounts of unaligned faces and can generalize well on test images without specific alignment. MyStyle [6], which relies on a personalized generative prior, shows very good identity retention, however the restored images lack fidelity to the input image. Blind restoration techniques, such as CodeFormer, GFP-GAN and DiffBIR lead to significant identity drifts. Compared to DiffBIR, our method achieves better fidelity to the ground truth face.

We also report quantitative results on our CelebRef-HQ test set in Table I. We report PSNR and SSIM as full-reference metrics, in addition to LPIPS and FID score. To quantify the identity disparity from the ground truth image, we use ArcFace [61] similarity. ASFFNet and DMDNet cannot restore the images effectively, resulting in low scores on all the metrics. Same is the case with DPS [58] and DiffFace [59]: both perform poorly across both reconstruction fidelity and identity retention metrics. MyStyle results in poor quantitative performance, even though the qualitative results show restored images with high fidelity. This can be attributed to the lack of faithfulness to the input image. Blind restoration methods (GFP-GAN, CodeFormer, DR2 and DiffBIR) show varying degrees of performance, however DiffBIR is found to perform

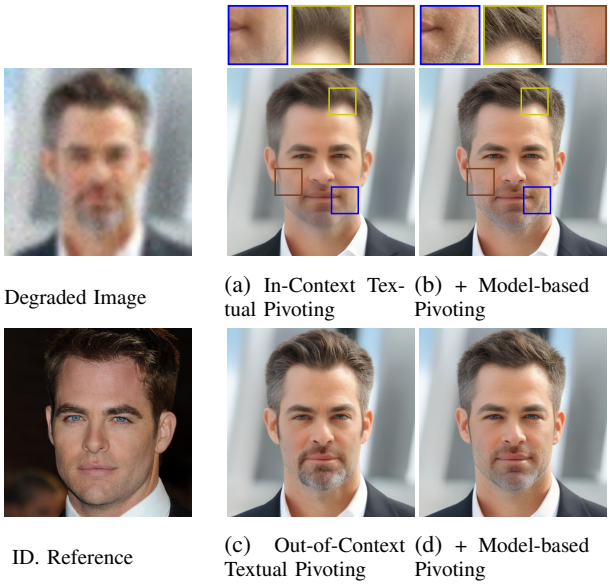
(a) DEGRADED IMAGE (b) ASFFNET [60] (c) DMDNET [5] (d) DIFFBIR [1] (e) OURS (f) ID. REFERENCE

Fig. 8: **Results on real degraded images.** Even in images in the wild with real, unknown degradation kernels, our proposed method is superior to the baselines in terms of identity preservation and perceptual quality.

the best among these. Our method achieves higher identity preservation, at the cost of slightly lower PSNR and SSIM to DiffBIR. While DiffBIR tends to generate smooth restorations to which PSNR and SSIM are not sensitive, our method injects realistic high-frequency details that may not be perfectly aligned with the ground truth image, thus decreasing the scores. Therefore, though PSNR and SSIM may be misleading metrics for the task at hand [62], the proposed method is still able to provide comparable quantitative fidelity, while retaining identity better. That being said, LPIPS and FID score, being metrics that favor high-frequency details, emphasize overall superiority of the proposed method. Specifically, across FID, LPIPS and ArcFace metrics, the proposed approach is state-of-the-art, both in terms of reconstruction fidelity as well as in terms of identity preservation.

b) User study.: We also conduct a user study (refer to Appendix A for detailed analysis). First, we analyze the benefit of our personalized prior towards perceived identity retention while being faithful to the input image. The proposed method is predominantly found to be superior, when compared with unconditional diffusion-based image restoration [1], reference-based image restoration [22], and GAN-based personalization [6]. Second, we surprisingly note that our personalized restoration method is found to improve identity-independent general image quality as well, when compared to the non-personalized diffusion-based restoration [1].

c) Comparison on real degradations.: Fig. 8 (and additional results in the supplement) show the performance of each method on real-world images. For each test image, the model is personalized with 10 reference images of the same identity from CelebRef-HQ. We also provide a reference image in column (f) for better evaluation of the identity preservation. This set of experiments show that our personalized model generalizes well to real-world degraded images. In all cases, our method achieves significantly better image quality than ASFFNet DMDNet, GFP-GAN and CodeFormer, and preserves identity better than DiFFBIR. When compared with MyStyle, in this operating regime the proposed method is considerably more faithful to the input degraded image, while being able to retain identity. **Please find more qualitative results in the supplement.**



Degraded Image (a) In-Context Textual Pivoting (b) + Model-based Pivoting

ID. Reference (c) Out-of-Context Textual Pivoting (d) + Model-based Pivoting

Fig. 9: **Ablation: understanding dual-pivot tuning.** (a) In-context textual pivoting alone enables a certain level of identity injection, although the details can be further improved. (b) Model-based pivoting improves the identity and perceptual quality (our method). (c) In contrast, out-of-context textual pivoting causes noticeable identity drifts. (d) Applying model-based pivoting in this setting also improves the results. The effects is much more pronounced, bridging the performance gap with the top row.

d) Training time analysis: Using the speedup strategies proposed in Section IV-C, we find that both textual and model-based pivoting steps see a twofold improvement in personalization time. In addition, we find that the model pivoting step can be extended to other identities in a zero-shot manner,

TABLE II: **Personalization speedup.** Both textual and model-based pivoting show speedup capabilities.

Step	Runtime (w/out speedup)	Runtime (w/ speedup)
Textual pivoting	63.5 min	31.8 min
Model pivoting	21.6 min	10.5 min

TABLE III: **Inference time.** The proposed method has comparable inference time to existing diffusion-based methods, with the added benefits of personalization.

Method	DifFace [59]	DR2 [28]	DiffBIR [1]	Ours
Runtime (s)	4.3	0.7	6.8	6.8

saving personalization time (Section IV-C). We note that the training is only once per subject: inference time per image is much faster, as shown in Table III, showing comparable runtimes as other diffusion-based methods while allowing strong personalization. While this is already reasonable for applications such as photo album restoration, improvements in these aspects will be beneficial: while this first work focuses on quality, follow-up works may focus on efficiency.

e) Ablation: Understanding Dual-Pivot Tuning: We find the proposed dual-pivot tuning approach to be a general technique for personalization of guided diffusion models. Fig. 9 qualitatively explores the benefits of this method. Specifically in our case, we find that alternate strategies for personalization exist. The first is our chosen strategy: in-context textual pivoting (Fig. 9(a)), which injects identity information, followed by model-based pivoting (Fig. 9(b)), which enables better utilization of the general restoration prior to get high-fidelity restored images. The alternate approach involves beginning with out-of-context textual pivoting (where $\mathbf{G}$ is personalized outside the context of $\mathbf{B}$). As shown in Fig. 9(c), this leads to significant gap in identity. However, post model-based pivoting (Fig. 9(d)), this is resolved, leading to high fidelity, identity preserving image restoration. The dual-pivot tuning approach successfully personalized the diffusion model in both these settings. We find the in-context textual pivoting to enable faster personalization in terms of the the model-based pivoting step.

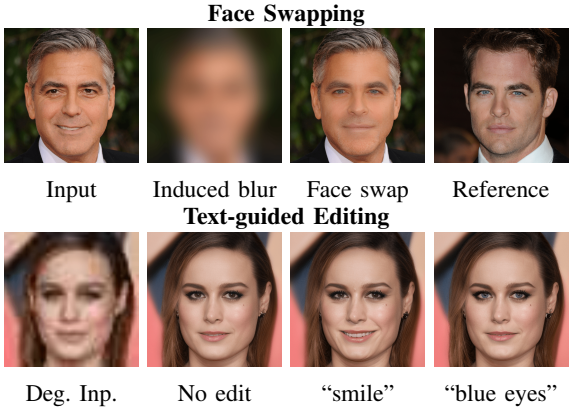


Fig. 10: **Editing applications: face swapping text-guided editing.** The upper row shows the application of face swapping. The swapped image shows structural similarity to the input (background, hair pose expression), while identity cues are drawn from the reference. The lower row shows performance on text-guided editing. We show two examples: enhancing the smile with the prompt "smile", and changing the eye color, with the prompt "blue eyes".

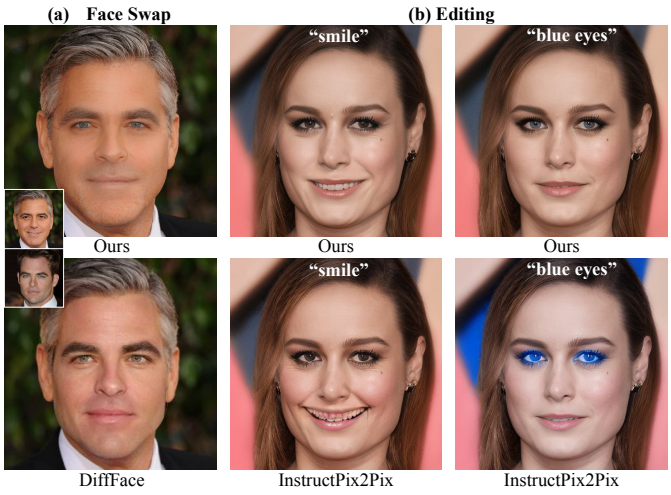


Fig. 11: **Comparing editing performance against existing methods.** We compare the face swap task against DiffFace [59], while we compare the text-guided editing task against InstructPix2Pix [52]. Across both the applications, we can see that the proposed method is qualitatively comparable or better than existing tailor-made methods for these tasks.

f) Additional Applications: Fig. 10 shows applications such as face swapping (top row) and text-guided editing (bottom row). We blur the source image to obscure identity. Then, for face swapping we restore using the personalized prior for a new identity. In the text-guided editing application we utilize the semantic prior of the diffusion model for text-conditional editing. For example, with the term "smiling" as part of the prompt, we see a smile in the expression of the individual in the output image. In addition, with the phrase "blue eyes", we see a clear variation in the eye color.

We also compare performance of editing tasks against specialized baselines for said tasks. While quantitative comparison is challenging, we show a qualitative analysis through Figure 11. For the face swap task, we compare with DiffFace [59]. As shown in Figure 11(a), the proposed method is comparable in terms of image quality swapped for face, while being more faithful to the input image (Figure 10 left), as well as to to the swapped identity (Figure 10 right). Additionally, identifying details such as eye color are better retained in our method.

For the text-guided image editing task, we compare with InstructPix2Pix [52]. Note that [52] is unable to operate on degraded images: we first use our method to restore the image (without any text editing prompt), which is then used as input to [52]. In contrast, our method applies directly to degraded images, and concurretly restores and edits the image. As can be seen, for both the prompts, our method leads to more realistic results, while enabling accurate editing control and remaining faithful to the input image.

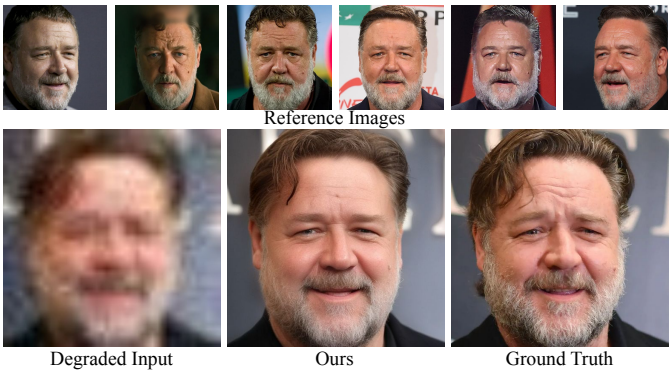


Fig. 12: **A visual description of limitations of the proposed method.** Since most of the reference training images for this particular identity do not show teeth, the restored image has a strong inductive bias towards not having teeth, even though the ground truth has clearly visible upper teeth.

VI. Conclusion

We propose a technique for personalizing a diffusion prior for face image restoration, leveraging the capabilities of few-shot fine-tuning based on a set of example images. Our method achieves high fidelity to both the input image and the identity of the person. We conduct extensive experiments to demonstrate the superiority of our method in comparison to various state-of-the-art alternatives.

a) Limitations and Future Work: While we make the first step in using personalized diffusion priors for face image restoration, despite the proposed speed-up techniques proposed for dual-pivot tuning, it remains a computationally expensive fine-tuning process for each identity. An important research direction is to inject few-shot identity in a feed-forward way, which we leave for future work. Further, while our personalized model improves fidelity to the identity, general fidelity and quality are fundamentally limited by the underlying restoration method. Additionally, potential limitations native to personalization [2] on a small set of images can find their way to our method, like overfitting towards features such as open eyes and smiling mouth if most of the training images show this and the input degraded image is ambiguous in this aspect (Figure 12 highlights this for the example where reference images do not show teeth and therefore the image restoration also removes teeth). Even though improving general quality is not the focus of this work, we believe contextual personalization holds promise towards improved performance in both image quality and identity fidelity.

Acknowledgments

The authors would like to thank Jackson Wang and Fangzhou Mu for helpful discussions and feedback on this manuscript.

Appendix A
Additional Quantitative Analysis

a) Additional notes on Tab. 1: An interesting result to note is the performance of [28]. We note that DR2 allows for tunable parameters (N, τ) selection to tradeoff restoration with fidelity to the original image. We choose $(N, \tau) = (16, 35)$ as we find this combination able to restore the degradations in our synthetic images. However, this does come at a cost of fidelity to the input - this is reflected by low fidelity as well as identity metrics for DR2.

(a) Inp. (b) DPS (c) DiffFace (d) Ours (e) GT

Fig. 13: **Additional qualitative comparisons against DiffFace [59] and Diffusion Posterior Sampling [58].** The proposed method is qualitatively superior across all examples, both in terms of image fidelity, as well as identity retention. Note aspects such as the eye color in row 1, and the mole on the upper left cheek in row 3, which is only retained through our method. Zoom for clearer comparison.

Another interesting observation is the performance of MyStyle [6] in terms of identity retention. We see that the ArcFace metric for MyStyle is similar to other blind restoration techniques. This is a somewhat surprising observation, since MyStyle uses a personalized prior for restoration. This may be explained as follows: while MyStyle indeed shows strong identity retention (as evidenced by qualitative results in Fig. 18, 20, 22, 23 in the supplementary material), the method is unable to retain fidelity (in terms of color, lighting, texture, makeup and even pose). All these factors can affect the perceived identity of the image from the perspective of the ArcFace metric, leading the this anomaly. On the other hand, the ArcFace metric is a valid comparison of identity retention with all other comparison methods, since they are able to retain fidelity with the input degraded image.

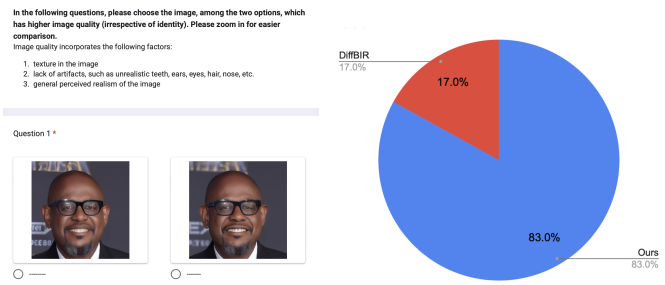
b) Additional Qualitative Baselines: Figure 13 shows two additional qualitative baselines: diffusion posterior sampling [58], a blind restoration method, and DiffFace [59], a face swapping-based method. We find that that proposed method is significantly superior, qualitatively, over these methods, both in terms of fidelity of restoration and identity retention. Both comparison methods particularly struggle with heavy degradations (rows 1 and 3), where identifying features such as eye color (row 1) and upper left cheek mole (row 3) are only maintained by our proposed method.

c) User study: We conduct a user study as a measure of perceptual quality and comparison with prior methods. We focus on two questions as part of the study: (a) can personalization help improve perceived image quality as well?,

and (b) how does our method compare with prior methods in terms of identity-aware restoration. We perform our study across 30 participants. The survey consists of randomized, anonymized options that they can choose amongst. Figs. 14 and 16 (a) show the exact guidelines provided as part of the study. For part (a), image quality, we compare against the base unconditional restoration method (DiffBIR [1]), with the objective being to choose the better-quality image, irrespective of identity. For this case, we compare across 11 image pairs, containing both real and synthetic degradations. For part (b), identity-aware restoration, we compare with three methods: (i) DMDNet [5] (better-performing method both qualitatively and quantitatively when compared with ASFFNet [60]); (ii) MyStyle [6], a personalized generative prior; (iii) DiffBIR [1], a unconditional diffusion-based method. For this case, we compare across 10 image sets, consists of synthetically degraded images.

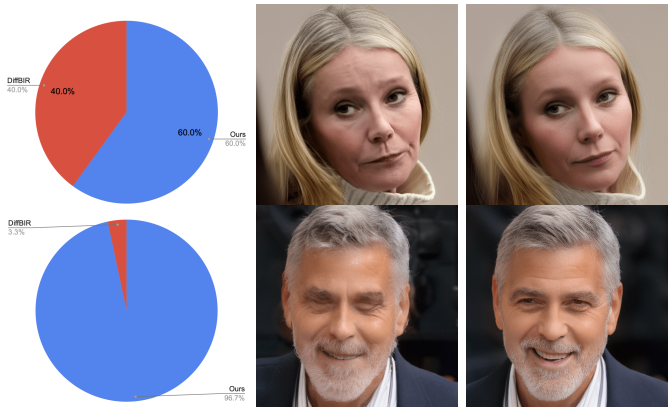
We begin with discussing the first part of the user study: effect of personalization on identity-independent perceived image quality. Fig. 14 shows the results of this part of the study. We observe that across the 30 study participants, a clear majority of the participants indicate the quality improvement that arises out of personalization. This is an unexpected result, and we explore this further by analyzing two specific instances, in Fig. 15. The upper row shows a case where the users are broadly split between the candidate options. The unconditional method provides a restored image with detailed facial features such as wrinkles, while the proposed method provides less of such detail, with stronger identity cues. Both images look viable as natural images with good quality. However, the lower row in Fig. 15 shows an instance where the study participants almost unanimously prefer our method. The reason for this is evident: the baselines method has considerable artifacts, especially near the eyes, while our method leads to a good quality, realistic face image. Through these observations, we can understand the effect of our personalization method on identity-independent image quality. In cases where the unconditional comparison method is able to perform, our personalized model remains stable and provides realistic looking faces. However, in cases where the comparison method fails, such as with high degradations, our method, throgh the strength of the identity prior, still results in realistic restored images. A combination of these two factors leads to superior perceived identity-independent image quality.

We next analyze the second part of the user study: the perceived strength of our identity-aware image restoration. Fig. 16 shows the results of this part of the study. We see that, perhaps per expectation, participants rate our method as the predominant favorite in terms of identity-aware image restoration, while retaining faithfulness to the input degraded image. This can be seen across our various qualitative results and speaks to the strength and reliability of our personalized prior across identities and degradations. However, an interesting insight is the relative placement of the comparison methods. Specifically, study participants rate DiffBIR [1] to be the second-best method, despite not retaining identity, as a result of its strong correlation to the input degraded image. On the other hand, MyStyle [6], while having a strong identity



(a) Example survey question (b) Survey Result

Fig. 14: **User study: effect of personalization on perceived image quality.** When asked to choose the image with better perceived quality, we find users predominantly choosing the images with our personalized restoration. Our method is indicated in blue, while DiffBIR [1] is shown in red, in the pie chart.



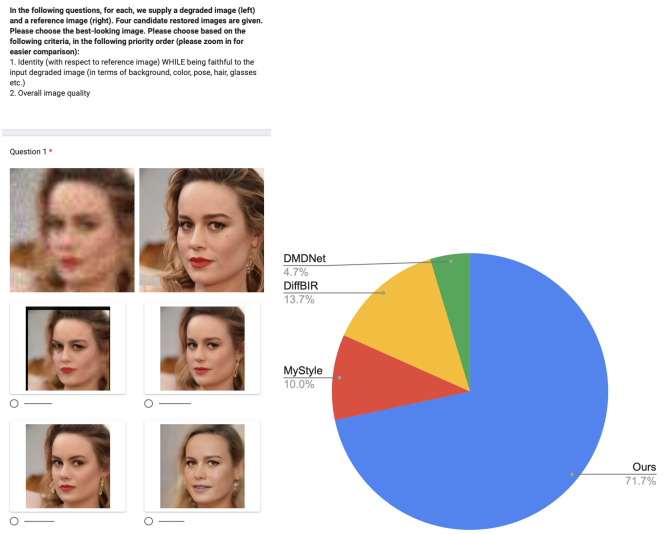
(a) Survey Result (quality) (b) DiffBIR [1] (c) Ours

Fig. 15: **Two specific real degradation restorations, from the lens of image quality.** The upper row shows a case where respondent opinion is split - DiffBIR provides specific detail like wrinkles, while our approach provides structure and identity. The lower row shows a case with unanimous favor towards our method. This arises from specific artifacts in the DiffBIR output, which is avoided by having a strong personalized prior. Our method is indicated in blue, while DiffBIR [1] is shown in red, in the pie charts.

prior, is the third-best preferred method on average, as a result of it not being faithful to the input image. That is, perceptually, faithfulness to the input degraded image is given a higher priority by participants, despite the study guidelines placing both identity and faithfulness to input image at the same priority.

APPENDIX B
ADDITIONAL OBSERVATIONS

a) Dealing with Heavy Degradations: Fig. 17 shows a potentially interesting use setting for the proposed method. Namely, in the case of very heavy degradations, multiple passes through the restoration model may be performed. As seen in the figure, (b) is able to obtain coarse details as well as



(a) Example survey question (b) Survey Result

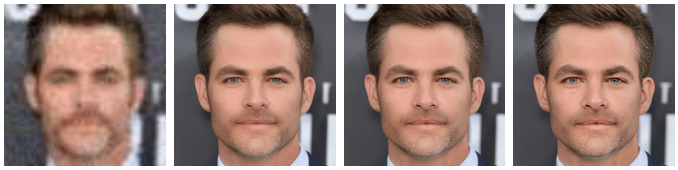
Fig. 16: **User study: the effect of personalization on perceived image identity retention.** When asked to choose the image with the better-looking image, while prioritizing identity and faithfulness to input degraded image, we find users predominantly choosing the images with our personalized restoration. Our method is indicated in blue, DiffBIR [1] is shown in yellow, MyStyle [6] is shown in red, and DMD-Net [5] is shown in green in the pie chart.

(a) (b) (c) (d)

Fig. 17: **Heavy degradation restoration.** Given a highly degraded image (a), a first pass (b) through our restoration pipeline provides a coarse estimate that pulls the image closer to the input domain of the model. The second pass (c) injects texture and identity information, leading to a better restored estimate with respect to the ground truth (d).

shape and strcuture information. Through a second pass, (c) is able to improve of texutre and detail, in addtion to obtaining greater identity injection, leading to a better restored image than (b).

b) The Effect of Classifier Free Guidance: Fig. 18 shows an ablation study on the effect of classifier-free guidance scale. As can be seen, the parameter allows for trading off image sharpness with realism, allowing for pereference-based tuning. Specifically, we find that increasing the CFG scale makes the restored images sharper, at the cost of realism. We also note in the main paper that the proposed method does not require conditional inference across all denoising steps. That is, to enable personalized restoration, we find that even though initial denoising steps are carried out unconditionally, effective personalized restoration is possible.



(a) Deg. inp. (b) CFG=2.0 (c) CFG=4.0 (d) CFG=6.0

Fig. 18: **Effect of classifier free guidance (CFG).** Sweeping across CFG enables a fidelity-diversity tradeoff, also manifesting as a tradeoff between image sharpness with realism (higher the CFG, higher the sharpness). This provides a degree of user control.

In Table IV, we show quantitative results. We note that traditional fidelity metrics worsen as the CFG is increased. This is consistent with our other observations, where the unconditional method (DiffBIR) shows slightly better performance on these metrics. As we increase the CFG, we move farther from the unconditional model and therefore see these effects. In terms of identity, we see a small reduction with increased CFG. As we increase the CFG, the sharpness of the restored image increases and may lead it to look unrealistic. Overall, in terms of tradional metrics, a lower CFG values is optimal. However, visually, the CFG can serve as a useful control knob for restored image style and quality. That being said, since we rely on existing CFG methods, the proposed method inherits the known limitations of CFG. For example, a very high CFG may lead to reduced realism and consistency with the input image or style.

APPENDIX C
IMPLEMENTATION DETAILS

We use the DiffBIR [1] model as the base architecture. The generator $\mathbf{G}$ is a Stable Diffusion v2.1 [7] as in [1], while the encoder $\mathbf{E}$ is initialized from [1]. Training is done using the AdamW optimizer [63]. We use a batch size of 1. Textual pivoting is trained with a learning rate of $8e-6$ for 2500 steps, and model-based pivoting is trained with a learning rate of $5e-5$ for 800 steps. The personalized token used for textual pivoting is kept as 'sks'. All experiments are conducted using an NVIDA V100 GPU. Please check included code for further specific details.

APPENDIX D
ETHICAL CONSIDERATIONS

The use of generative models, while being extremely helpful for challenging tasks such as identity-aware image restoration, can also potentially have harmful effects. Specifically, generative models can be used for immoral tasks. In our case, applications that we discussed, such as face swapping as well as text-guided editing, can lead to generation of fake images, that may be used without the consent of the person whose image it is. We strongly condemn such and any other harmful use cases of the proposed method.

TABLE IV: **Ablation: effect of classifier-free guidance scale on personalized image restoration**. We compare along PSNR and SSIM as fidelity metrics, while ArcFace similarity serves as an identity metric.

CFG	PSNR (dB)	SSIM	ArcFace (Identity)
1.0	24.78	0.70	0.90
2.0	24.54	0.69	0.89
3.0	24.17	0.68	0.87
4.0	23.73	0.67	0.85
5.0	23.12	0.65	0.84

REFERENCES

[1] 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. 1, 2, 3, 4, 5, 6, 7, 8, 9, 11, 12

[2] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 500–22 510. 1, 2, 3, 5, 10

[3] Z. Wang, J. Chen, and S. C. Hoi, "Deep learning for image super-resolution: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 10, pp. 3365–3387, 2020. 1

[4] 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*, 2020, pp. 2706–2715. 1, 3, 7

[5] 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*, 2022. 1, 3, 6, 7, 8, 11, 12

[6] 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)*, vol. 41, no. 6, pp. 1–10, 2022. 1, 3, 7, 8, 10, 11, 12

[7] 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*, 2022, pp. 10 684–10 695. 1, 3, 4, 6, 12

[8] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *Advances in Neural Information Processing Systems*, vol. 35, pp. 36 479–36 494, 2022. 1, 3

[9] J. Wang, Z. Yue, S. Zhou, K. C. Chan, and C. C. Loy, "Exploiting diffusion prior for real-world image super-resolution," *arXiv preprint arXiv:2305.07015*, 2023. 1, 4

[10] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, 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. 1, 3

[11] Y. Chen, Y. Tai, X. Liu, C. Shen, and J. Yang, "Fsrnet: End-to-end learning face super-resolution with facial priors," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2492–2501. 3

[12] A. Bulat and G. Tzimiropoulos, "Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 109–117. 3

[13] D. Kim, M. Kim, G. Kwon, and D.-S. Kim, "Progressive face super-resolution via attention to facial landmark," *arXiv preprint arXiv:1908.08239*, 2019. 3

[14] 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*, 2021, pp. 11 896–11 905. 3

[15] Z. Shen, W.-S. Lai, T. Xu, J. Kautz, and M.-H. Yang, "Deep semantic face deblurring," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8260–8269. 3

[16] L. Yang, S. Wang, S. Ma, W. Gao, C. Liu, P. Wang, and P. Ren, "Hifacegan: Face renovation via collaborative suppression and replenishment," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 1551–1560. 3

[17] X. Yu, B. Fernando, B. Ghanem, F. Porikli, and R. Hartley, "Face super-resolution guided by facial component heatmaps," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 217–233. 3

[18] X. Li, C. Chen, S. Zhou, X. Lin, W. Zuo, and L. Zhang, "Blind face restoration via deep multi-scale component dictionaries," in *European Conference on Computer Vision*. Springer, 2020, pp. 399–415. 3

[19] Z. Wang, J. Zhang, R. Chen, W. Wang, and P. Luo, "Restoreformer: High-quality blind face restoration from undegraded key-value pairs," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 512–17 521. 3

[20] Y. Zhao, Y.-C. Su, C.-T. Chu, Y. Li, M. Renn, Y. Zhu, C. Chen, and X. Jia, "Rethinking deep face restoration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 7652–7661. 3

[21] 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," *arXiv preprint arXiv:2205.06803*, 2022. 3

[22] 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*, vol. 35, pp. 30 599–30 611, 2022. 3, 7, 8

[23] K. C. Chan, X. Wang, X. Xu, J. Gu, and C. C. Loy, "Glean: Generative latent bank for large-factor image super-resolution," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 14 245–14 254. 3

[24] T. Yang, P. Ren, X. Xie, and L. Zhang, "Gan prior embedded network for blind face restoration in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 672–681. 3

[25] X. Wang, Y. Li, H. Zhang, and Y. Shan, "Towards real-world blind face restoration with generative facial prior," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3, 7

[26] X. Luo, X. Zhang, P. Yoo, R. Martin-Brualla, J. Lawrence, and S. M. Seitz, "Time-travel rephotography," *ACM Transactions on Graphics (TOG)*, vol. 40, no. 6, pp. 1–12, 2021. 3

[27] Z. Yue and C. C. Loy, "Difface: Blind face restoration with diffused error contraction," *arXiv preprint arXiv:2212.06512*, 2022. 3

[28] Z. Wang, Z. Zhang, X. Zhang, H. Zheng, M. Zhou, Y. Zhang, and Y. Wang, "Dr2: Diffusion-based robust degradation remover for blind face restoration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1704–1713. 3, 7, 9, 10

[29] Y. Du, T. Hu, R. Yi, and L. Ma, "Ld-bfr: Vector-quantization-based face restoration model with latent diffusion enhancement," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 2852–2860. 3

[30] X. Li, M. Liu, Y. Ye, W. Zuo, L. Lin, and R. Yang, "Learning warped guidance for blind face restoration," in *The European Conference on Computer Vision (ECCV)*, September 2018. 3

[31] B. Dogan, S. Gu, and R. Timofte, "Exemplar guided face image super-resolution without facial landmarks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019. 3

[32] M. J. Chong, D. Xu, Y. Zhang, Z. Wang, D. Forsyth, G. Krishnan, Y. Wu, and J. Wang, "Copy or not? reference-based face image restoration with fine details," in *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, February 2025, pp. 9642–9651. 3

[33] K. Wang, J. Oramas, and T. Tuytelaars, "Multiple exemplars-based hallucination for face super-resolution and editing," in *Proceedings of the Asian Conference on Computer Vision*, 2020. 3

[34] L. Zeng, L. Chen, Y. Xu, and N. K. Kalantari, "Mystyle++: A controllable personalized generative prior," in *SIGGRAPH Asia 2023 Conference Papers*, 2023, pp. 1–11. 3

[35] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020. 3

[36] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=PxTIG12RRHS 3

[37] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021. 3

[38] 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. 3

[39] R. Po, W. Yifan, V. Golyanik, K. Aberman, J. T. Barron, A. H. Bermano, E. R. Chan, T. Dekel, A. Holynski, A. Kanazawa *et al.*, "State of the art on diffusion models for visual computing," *arXiv preprint arXiv:2310.07204*, 2023. 3

[40] W. Chen, H. Hu, Y. Li, N. Ruiz, X. Jia, M.-W. Chang, and W. W. Cohen, "Subject-driven text-to-image generation via apprenticeship learning," *arXiv preprint arXiv:2304.00186*, 2023. 3

[41] Y. Tewel, R. Gal, G. Chechik, and Y. Atzmon, "Key-locked rank one editing for text-to-image personalization," in *ACM SIGGRAPH 2023 Conference Proceedings*, 2023, pp. 1–11. 3

[42] L. Tang, N. Ruiz, Q. Chu, Y. Li, A. Holynski, D. E. Jacobs, B. Hariharan, Y. Pritch, N. Wadhwa, K. Aberman *et al.*, "Realfill: Reference-driven generation for authentic image completion," *arXiv preprint arXiv:2309.16668*, 2023. 3

[43] Y. Alaluf, E. Richardson, G. Metzer, and D. Cohen-Or, "A neural space-time representation for text-to-image personalization," *arXiv preprint arXiv:2305.15391*, 2023. 3

[44] A. Voynov, Q. Chu, D. Cohen-Or, and K. Aberman, "$p+$: Extended textual conditioning in text-to-image generation," *arXiv preprint arXiv:2303.09522*, 2023. 3, 5

[45] Z. Liu, R. Feng, K. Zhu, Y. Zhang, K. Zheng, Y. Liu, D. Zhao, J. Zhou, and Y. Cao, "Cones: Concept neurons in diffusion models for customized generation," *arXiv preprint arXiv:2303.05125*, 2023. 3

[46] R. Gal, M. Arar, Y. Atzmon, A. H. Bermano, G. Chechik, and D. Cohen-Or, "Encoder-based domain tuning for fast personalization of text-to-image models," *ACM Transactions on Graphics (TOG)*, vol. 42, no. 4, pp. 1–13, 2023. 3

[47] M. Arar, R. Gal, Y. Atzmon, G. Chechik, D. Cohen-Or, A. Shamir, and A. H. Bermano, "Domain-agnostic tuning-encoder for fast personalization of text-to-image models," *arXiv preprint arXiv:2307.06925*, 2023. 3

[48] N. Ruiz, Y. Li, V. Jampani, W. Wei, T. Hou, Y. Pritch, N. Wadhwa, M. Rubinstein, and K. Aberman, "Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models," *arXiv preprint arXiv:2307.06949*, 2023. 3

[49] R. Po, G. Yang, K. Aberman, and G. Wetzstein, "Orthogonal adaptation for modular customization of diffusion models," *arXiv preprint arXiv:2312.02432*, 2023. 3

[50] O. Avrahami, K. Aberman, O. Fried, D. Cohen-Or, and D. Lischinski, "Break-a-scene: Extracting multiple concepts from a single image," *arXiv preprint arXiv:2305.16311*, 2023. 3

[51] B. Kawar, S. Zada, O. Lang, O. Tov, H. Chang, T. Dekel, I. Mosseri, and M. Irani, "Imagic: Text-based real image editing with diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6007–6017. 3

[52] T. Brooks, A. Holynski, and A. A. Efros, "Instructpix2pix: Learning to follow image editing instructions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 392–18 402. 3, 9

[53] S. Wang, C. Saharia, C. Montgomery, J. Pont-Tuset, S. Noy, S. Pellegrini, Y. Onoe, S. Laszlo, D. J. Fleet, R. Soricut *et al.*, "Imagen editor and editbench: Advancing and evaluating text-guided image inpainting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 359–18 369. 3

[54] 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*, 2023, pp. 3836–3847. 3, 4

[55] C. Mou, X. Wang, L. Xie, Y. Wu, J. Zhang, Z. Qi, Y. Shan, and X. Qie, "T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models," *arXiv preprint arXiv:2302.08453*, 2023. 3

[56] J. Ho and T. Salimans, "Classifier-free diffusion guidance," *arXiv preprint arXiv:2207.12598*, 2022. 3

[57] D. Roich, R. Mokady, A. H. Bermano, and D. Cohen-Or, "Pivotal tuning for latent-based editing of real images," *ACM Transactions on graphics (TOG)*, vol. 42, no. 1, pp. 1–13, 2022. 4

[58] H. Chung, J. Kim, M. T. Mccann, M. L. Klasky, and J. C. Ye, "Diffusion posterior sampling for general noisy inverse problems," *arXiv preprint arXiv:2209.14687*, 2022. 7, 10

[59] K. Kim, Y. Kim, S. Cho, J. Seo, J. Nam, K. Lee, S. Kim, and K. Lee, "Diffface: Diffusion-based face swapping with facial guidance," *Pattern Recognition*, p. 111451, 2025. 7, 9, 10

[60] 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 *CVPR*, 2020. 7, 8, 11

[61] 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*, 2019, pp. 4690–4699. 7

[62] J. Gu, H. Cai, C. Dong, J. S. Ren, R. Timofte, Y. Gong, S. Lao, S. Shi, J. Wang, S. Yang *et al.*, "Ntire 2022 challenge on perceptual image quality assessment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 951–967. 8

[63] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017. 12