AdaFace – A Versatile Face Encoder for Zero-Shot Diffusion Model Personalization

Anonymous Author(s)
Affiliation
Address
email

Figure 1: Although AdaFace is solely trained on static images, the subject embeddings it generates can directly condition AnimateDiff to produce personalized videos across diverse scenes without requiring any modifications.

Abstract

Since the advent of diffusion models, personalizing these models – conditioning them to render novel subjects – has been widely studied. Recently, several methods propose training a dedicated image encoder on a large variety of subject images. This encoder maps the images to identity embeddings (ID embeddings). During inference, these ID embeddings, combined with conventional prompts, condition a diffusion model to generate new images of the subject. However, such methods often face challenges in achieving a good balance between authenticity and compositionality – accurately capturing the subject's likeness while effectively integrating them into varied and complex scenes. A primary source for this issue is that the ID embeddings reside in the *image token space* ("image prompts"), which is not fully composable with the text prompt encoded by the CLIP text encoder. In this work, we present AdaFace, an image encoder that maps human faces into the *text prompt space*. After being trained only on 400K face images with 2 GPUs, it achieves high authenticity of the generated subjects and high compositionality with various text prompts. In addition, as the ID embeddings are integrated in a normal text prompt, it is highly compatible with existing pipelines and can be used without modification to generate authentic videos. We showcase the generated images and videos of celebrities under various compositional prompts. The source code is released on an anonymous repository `https://github.com/adaface-neurips/adaface`.

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

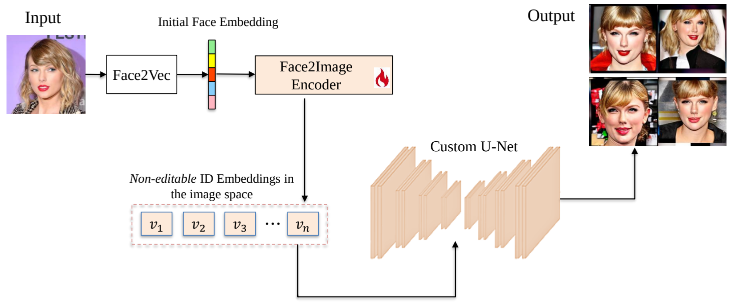

Figure 2: A typical zero-shot face encoder pipeline for diffusion models. First, a Face2Vec module (e.g., ArcFace [Deng et al., 2019]) extracts a single vector that captures the facial features. Then a trainable Face2Image encoder (e.g., Arc2Face [Papantoniou et al., 2024]) maps it to n facial tokens $v_1, \cdots, v_n$ within the image embedding spaces. The facial tokens condition the U-Net (either original or fine-tuned) to generate authentic-looking subject images. However, since the facial tokens is not blended with other text prompts (sometimes they are simply concatenated), the whole pipeline has weaker compositionality than using text prompts alone. Moreover, such models are often incompatible with existing diffusion pipelines, such as AnimateDiff Guo et al. [2024a].

1 Introduction

Recent years have witnessed the blossom of diffusion models, which have been widely used in image generation, image editing, and video generation [Ho et al., 2020, Nichol et al., 2022, Saharia et al., 2022, Rombach et al., 2022, Podell et al., 2024, Chen et al., 2024a, Kawar et al., 2023, Peebles and Xie, 2023, Guo et al., 2024a]. A particularly interesting application of these models is personalization, where they are conditioned to generate images of specific subjects. Previously, this was primarily achieved through test-time fine-tuning [Ruiz et al., 2022, Gal et al., 2022a, Kumari et al., 2022, Tewel et al., 2023], which introduced additional computational demands and complexity to the image generation process. Recent advancements have seen the development of zero-shot, tuning-free methods [Wei et al., 2023, Ye et al., 2023, Shi et al., 2023, Wang et al., 2024, Papantoniou et al., 2024, Guo et al., 2024b, Huang et al., 2024, Han et al., 2024, Chen et al., 2024b, He et al., 2024]. These methods train a dedicated image encoder to convert subject images to identity embeddings (ID embeddings) using a large dataset. During inference, these ID embeddings are combined with standard text prompts to generate new images of the subject (Figure 2). Despite these innovations, these approaches often struggle to strike a good balance between authenticity and compositionality. Authenticity ensures the model captures the true likeness of the subject, whereas compositionality concerns the model's ability to seamlessly integrate the subject into diverse and intricate scenes. The challenge primarily stems from how ID embeddings are utilized: in many zero-shot methods, the embeddings exist in the *image token space* ("image prompts") and do not fully mesh with text prompts. In cases like [Huang et al., 2024], while the ID embeddings are within the text space, there lacks targeted training to enhance their integration with other text prompts, resulting in compromised compositionality.

Given the limitations of existing methods, we propose AdaFace, a versatile face encoder that maps human faces into the *text prompt space*. First, the ID embeddings generated by AdaFace seamlessly integrate with text prompts via the CLIP text encoder, allowing for more coherent and expressive conditioning. Second, we employ targeted training strategies to enhance the compositionality of the ID embeddings, ensuring they are able to be used to generate diverse and complex scenes. Furthermore, AdaFace is highly compatible with existing diffusion pipelines, requiring no modifications to generate authentic videos, as demonstrated in Figure 1. Notably, due to efficient model design and distillation techniques, AdaFace is trained on merely 406,567 face images with 2 RTX A6000 GPUs, all within a constrained compute budget.

We demonstrate the effectiveness of AdaFace by showcasing the generated images and videos of celebrities under various compositional prompts. We also perform quantitative evaluations to validate that AdaFace achieves a good balance between authenticity and compositionality, measured by ArcFace similarity and CLIP-Text similarity, respectively.

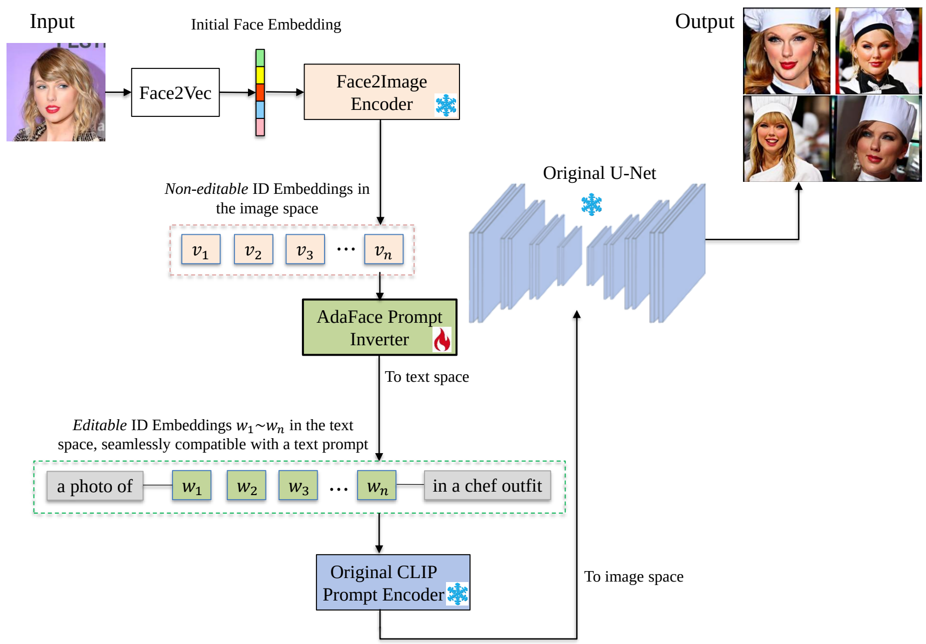

Figure 3: The core of AdaFace is the *Prompt Inverter*, which inverts the image-space ID embeddings from another model to the *text prompt space*, represented as $w_1, \cdots, w_n$. These embeddings are integrated into a standard text prompt and encoded by a CLIP prompt encoder. CLIP coherently composes the semantics of the ID embeddings and the text prompt, providing good compositionality.

2 Method

Motivated by the advantages of *text space* face prompts, we propose techniques to distill one or more image-space face encoders into the text space, and further enhance its compositionality. The overall architecture of AdaFace is shown in Figure 3. The core module of AdaFace is the *AdaFace Prompt Inverter*, which inverts the image-space ID embeddings to the text space, enabling the integration of the ID embeddings into a standard text prompt. The ID embeddings are then encoded by a CLIP prompt encoder, which coherently composes the semantics of the ID embeddings and the text prompt. The text-level composition also facilitates *Composition Distillation* (Figure 5), which significantly improves the compositionality of the ID embeddings without additional training data. A side-effect of composition distillation is that, when there is spatial misalignment between the subject-single and subject-composition images, the subject features will be gradually contaminated by background features, reducing their authenticity. Accordingly, we propose a *Elastic Face Preserving Loss* (Figure 6), to prevent the subject features from degeneration.

2.1 AdaFace Architecture

The core module of AdaFace is the *AdaFace Prompt Inverter*, which converts the image-space ID embeddings from a Face2Image model to the text space.

The architecture and initialization of the prompt inverter significantly impacts the training efficiency. Compared to other deep learning tasks, the diffusion training is highly stochastic and the gradients have a much lower signal-to-noise ratio. It is highly challenging to train a sizable diffusion component from scratch without high compute budgets and large batch sizes. To achieve efficient learning, we adopt the same architecture as the CLIP text encoder for the AdaFace Prompt Inverter, and initialize it with the pre-trained weights. This ensures that the output embeddings are not very distant from the text space from the beginning of training, and the model learns more signals from the gradients.

One may raise the question that since the output of a pre-trained CLIP encoder is in the image space, why it is able to adapt quickly to generate text-space embeddings? We speculate that in CLIP, the semantics of low-level layers and high-level layers are not in totally incompatible spaces, but rather,

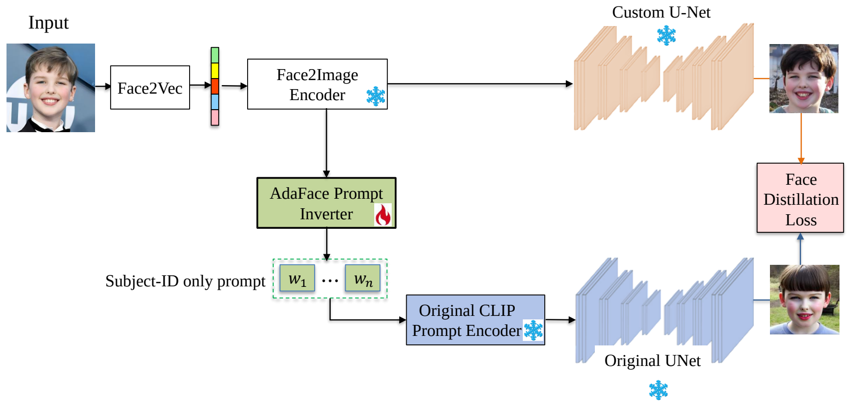

Figure 4: Face distillation on face images. The output of the AdaFace stream is compared with the Face2Image stream. During this process, only the AdaFace Prompt Inverter is optimized.

the high-level semantics enrich the low-level ones. Our hypothesis is corroborated by [Toker et al., 2024], as well as the community practice of ad-hoc fusing the output embeddings of multiple CLIP text encoder layers[1]. The semantics of layer features gradually transition from the text space to the image space. As a result, during fine-tuning, the skip connections within CLIP will allow the low-level semantics to take shortcut towards the output embeddings, and the high-level layers will gradually learn to enrich the low-level semantics in the *text space* instead.

The training of the prompt inverter is divided into two stages. In the first *face distillation* stage, a Face2Image model guides the prompt inverter to generate authentic faces in the text prompt space. In the second *composition distillation* stage, the prompt inverter observes how the original model output responds to compositional prompts, and learns to generate similar responses, so as to allow the text prompts to control the composition of the generated images.

2.2 Face Distillation

The face distillation stage is illustrated in Figure 4, where the objective is to minimize the difference between the generated images by the original Face2Image model and by the AdaFace Prompt Inverter on the same initial noise. The training objective, namely the face distillation loss, is formulated as a reconstruction loss between the two generated images:

$$\mathcal{L}_{\text{face}} = \mathbb{E}_{f\sim F, z\sim\mathcal{N}(0,I), t\in[1,T]}\left[\|G_{\text{AdaFace}}(f,z,t|\theta) - G_{\text{Face2Image}}(f,z,t|\theta')\|_2^2\right], \quad (1)$$

where $G_{\text{Face2Image}}$ and G_{AdaFace} are the Face2Image and the AdaFace Prompt Inverter conditioned U-Nets, respectively, f is a random face drawn from the face space F, z is the initial noise, and θ and θ' are the parameters of the AdaFace Prompt Inverter and the Face2Image model, respectively. For some models such as Ada2Face, $\theta' \neq \theta$.

In order to sweep the input space $\{f, z, t\}$ as completely as possible, we adopt a few techniques:

Random Gaussian Face Embeddings. Empirically, we observe that almost all random face embeddings result in legitimate face images when processed by the Face2Image model. Therefore, we expand the candidate face space F by including random face embeddings drawn from a Gaussian distribution, alongside the face embeddings extracted from real face images: $F = F_{\text{real}} \cup F_{\text{rand}}$.

Multi-Timestep Distillation. We use multiple denoising steps on the same initial noise, and compute the reconstruction loss on all the steps, so that the prompt inverter learns to imitate the Face2Image model's behavior on intermediate noise levels:

$$\mathcal{L}_{\text{face}} = \mathbb{E}_{f\sim F, z_1\sim\mathcal{N}(0,I), t_1>\cdots>t_k\in[1,T]}\sum_{i=1}^{k}\left[\|G_{\text{AdaFace}}(f,z_i,t_i|\theta) - G_{\text{Face2Image}}(f,z_i,t_i|\theta')\|_2^2\right], \quad (2)$$

[1]https://github.com/AUTOMATIC1111/stable-diffusion-webui/discussions/5674

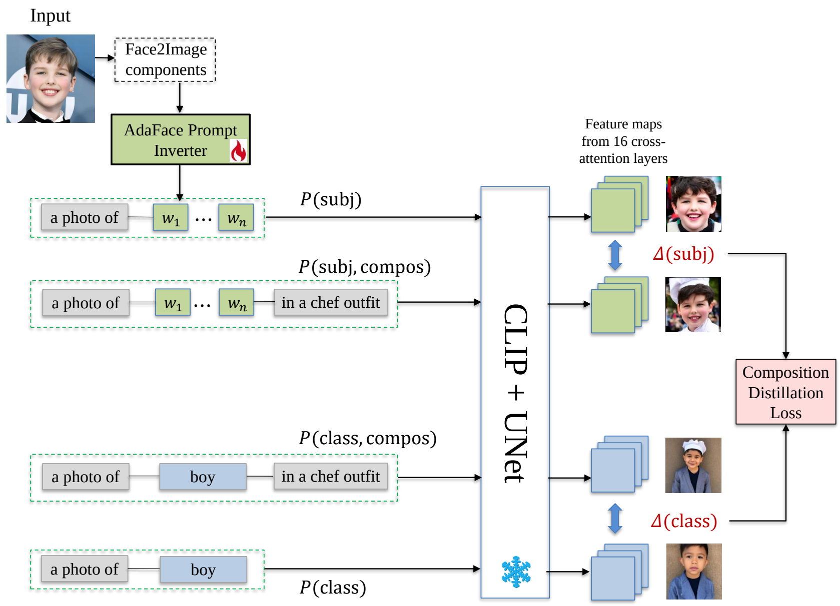

Figure 5: Composition distillation on four types of prompts: subject-single, subject-composition, class-single and class-composition. The four generated images form two contrastive pairs, and their feature deltas are encouraged to be aligned through a composition distillation loss.

where $t_1, \cdots, t_k$ are a randomly sampled sequence of timesteps, and when $i > 1$, z_i is the partially denoised image by $G_{\text{Face2Image}}$ in the previous step.

Dynamic Model Expansion. When the training loss plateaus, it suggests that the model has reached the limits of its capacity to capture nuanced facial features. In this situation, we expand the model capacity by incorporating additional *query* and *value* projections within the attention layers of the prompt inverter. As a result, each token is represented by multiple, subtly distinct query and value tokens. This enables the model to better grasp the subtle facial features of the subject, thanks to the increased diversity and richness of the queries and values. Note that the number of keys and output tokens remain unchanged, ensuring that the computational load does not increase drastically.

Specifically, when a query projection Q is expanded by N times, we make N identical copies of Q and add Gaussian noises to $N-1$ of them. The same operation is applied to the value projection V. This is to ensure that the expanded Q' and V' do not deviate too much from the original Q and V, and the model augments the original features with slightly varied replicas.

The attention expansion proves to be particularly beneficial at the lower layers of the prompt inverter. Intuitively, once some information in the features from the upstream Face2Image encoder is lost in the lower layers, it is hard to recover in the higher layers. The mechanism of expanding queries and values creates multiple, slightly varied replicas of the same information, thereby allowing the model to select the most informative copy for preservation and further processing in subsequent layers. This approach is conceptually akin to the role of the excitation operator in a squeeze-and-excitation network [Hu et al., 2018], which also emphasizes selectively retaining the most significant features.

2.3 Composition Distillation

A prevalent issue with existing face encoders is that the subject token tends to dominate the generated images, resulting in degeneration of compositionality. To mitigate this issue, we employ *composition distillation* (Figure 5) to regularize the subject embeddings, ensuring that their semantics are effectively integrated with other tokens, enhancing the overall expression. During this process, the model observes how the original diffusion model adjusts output features to incorporate additional compositional prompts into the output image. The model then imitates these adjustments when encountering similar compositional prompts.

For this purpose, four types of prompts are employed to form two contrastive pairs: 1) a "subject-single" prompt that only contains the subject, such as "A photo of a [Zendaya]", 2) a "subject-composition" prompt such as "A photo of a [Zendaya] in the forest", 3) a "class-single" prompt that only contains a general class, such as "A photo of a woman", and 4) a "class-composition" prompt such as "A photo of a woman in the forest". Ideally, the semantic differences between "A photo of x" and "A photo of x in the forest" should only be relevant to "in the forest", and is independent of x.

We represent the semantic differences between two pairs of prompts as their "feature deltas". The training objective is to encourage the feature deltas between the subject-single and subject-composition images to be aligned with the feature deltas between the class-single and class-composition images. In other words, the following equation is expected to hold approximately:

$$\begin{aligned} &\Delta(\text{subject, compos}) \doteq \text{feat}(\text{subject, compos}) - \text{feat}(\text{subject}) \\ \approx &\Delta(\text{class, compos}) \quad \doteq \text{feat}(\text{class, compos}) \quad - \text{feat}(\text{class}), \end{aligned} \tag{3}$$

where subject, class, (subject, compos) and (class, compos) denote the four types of prompts, respectively. (subject, compos) and (class, compos) are randomly drawn from a pool of common compositional prompts consisting of various backgrounds, additional objects, dresses, image styles and lighting conditions. $\text{feat}(x)$ refers to relevant features, including 1) the output features from all the cross-attention layers, 2) the attention maps in all the cross-attention layers, and 3) the encoded prompt embeddings by CLIP text encoder. $\text{feat}(x) - \text{feat}(y)$ is the *orthogonal subtraction* between two feature maps, defined below.

We define a *compositional delta loss* that aligns the feature deltas $\Delta_i(\text{subject, compos})$ and $\Delta_i(\text{class, compos})$ on the three types of features listed above:

$$\mathcal{L}_\Delta = \sum_i \{1 - \mathbb{E}_{\text{compos}\sim \text{U}(C)} \cos(\Delta_i(\text{subject, compos}), \Delta_i(\text{class, compos}))\}, \tag{4}$$

in which i indexes the feature type (cross-attention output features, attention maps or CLIP prompt embeddings), and $\text{U}(C)$ is a uniform distribution on a set of compositional prompts C.

Orthogonal Subtraction. We wish to remove subject-specific features through the feature subtraction "feat(subject, compos) $-$ feat(subject)". However, it is commonly observed that the subject-specific features may have different magnitudes (often smaller under compositional prompts). To mitigate this issue, we propose to use orthogonal subtraction, which is invariant to the scale of the subject-specific features. A relevant idea [Wang et al., 2023] is explored for language model fine-tuning. Specifically, the feature deltas are calculated using the following equation:

$$\Delta\text{feat}(s, c) = \text{feat}(s, c) - \text{proj}_{\text{feat}(s)}(\text{feat}(s, c)), \tag{5}$$

where $\text{proj}_{\text{feat}(s)}(\text{feat}(s, c))$ is the projection of $\text{feat}(s, c)$ onto $\text{feat}(s)$, computed as:

$$\text{proj}_{\text{feat}(s)}(\text{feat}(s, c)) = \langle \text{feat}(s, c), \text{feat}(s) \rangle \text{feat}(s), \tag{6}$$

with $\langle \text{feat}(s, c), \text{feat}(s) \rangle$ being the inner product between the two features. The operation effectively projects $\text{feat}(s, c)$ onto the orthogonal complement of $\text{feat}(s)$ and then subtracts this projection from $\text{feat}(s, c)$. As a result, $\Delta\text{feat}(s, c)$, the feature delta, is orthogonal to $\text{feat}(s)$. This methodology ensures that the deltas remove as much of the subject-specific features as possible, thereby minimizing the influence of the scales of the subject-specific features contained within $\text{feat}(s, c)$.

Differences with Previous Methods. While previous methods have explored analogous concepts, such as StyleGAN-NADA [Gal et al., 2022b], which applies similar regularizations in the CLIP prompt embedding space, and PuLID [Guo et al., 2024b], which introduces similar contrastive regularizations on cross-attention queries, our approach is more comprehensive and effective. Our compositional delta loss encompasses a broader range of relevant features, including the attention maps and output features from cross-attention layers, and the CLIP prompt embeddings. Moreover, we introduce an orthogonal subtraction technique for computing the feature deltas. This technique isolates and extracts composition-specific features, making the distillation more effective.

2.4 Elastic Face Preserving Loss

The composition distillation is done on instances with different prompts starting from the same initial noise. This is to encourage the diffusion model to generate images that are compositionally similar

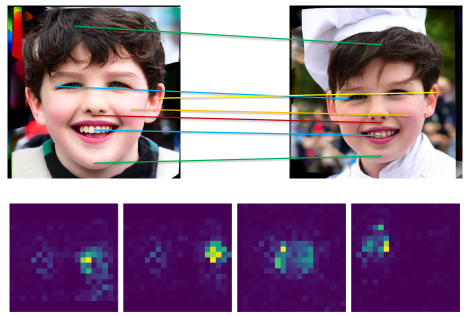

Figure 6: To prevent subject features from degeneration due to spatial misalignment during composition distillation, we propose a Elastic Face Preserving Loss. The second row shows the cross-attention maps at selected four points on the subject-single image. The highlighted pixels associate the corresponding facial areas across the two images. The features of matching pixels are required to be close to each other to achieve subject feature preservation.

[Zhang et al., 2024], to achieve more accurate alignment between the image pairs. Despite this effort, spatial misalignment often persists between the images differently prompted. This misalignment can result in delta loss providing erroneous signals from non-facial to facial areas, slowly reducing the authenticity of the generated subjects. For instance, on a noisy input face image, the output image from the subject-single instance is expected to largely retain the same facial contours as the input. However, the output from the subject-composition instance often deviate from the original face contours, due to the introduction of additional compositional elements. An illustrative example provided in the first row of Figure 6 shows how a chef hat in one image spatially aligns with the hair in another, leading to potential contamination in the subject's hair representations.

To tackle this challenge, we view the subject-composition image as a "warped" version of the subject-single image, and turn to techniques from the Optical Flow literature[Teed and Deng, 2020, Sui et al., 2022] to estimate a matching field. The matching field is used to spatially align the subject features across different images, ensuring them to be consistently maintained after "warping".

Specifically, the model takes as input a noisy face image from the training data. The face image is accompanied by a segmentation mask, isolating the face area for matching. We compute the cross attention matrix[2] between the queries of a subject-single instance and a subject-composition instance:

$$\mathrm{CA}(\mathrm{subj}, \mathrm{compos}) = \mathrm{softmax}(\mathrm{Q}_{\mathrm{subj}}\mathrm{Q}_{\mathrm{compos}}^{T}), \tag{7}$$

By looking up the cross-attention map $\mathrm{CA}(\mathrm{subj}, \mathrm{compos})$, we can find the pixels best matching a subject-single image pixel in a subject-composition image. The second row in Figure 6 shows the attention maps of four points on the face in the left image. We "soft-warp" the subject-composition features to align with the subject-single features through matrix multiplication, and require the warped features to be close to the facial features in the subject-single image:

$$\mathcal{L}_{\text{face-preserving}} = 1 - \cos\Big(\mathrm{CA}(\mathrm{subj}, \mathrm{compos}) \odot \mathrm{feat}(\mathrm{compos}), \mathrm{feat}(\mathrm{subj})\Big)_{\mathrm{mask}}. \tag{8}$$

Here for clarity, $\mathrm{feat}(\mathrm{subject}, \mathrm{compos})$ is abbreviated as $\mathrm{feat}(\mathrm{compos})$. The cosine similarity $\cos(\cdot, \cdot)$ is computed on the masked area. The face-preserving loss is computed on each U-Net cross-attention layer. It encourages the subject features in the subject-composition instance to be consistent with those in the subject-single instance, preventing them from being contaminated in the composition distillation process.

[2]The inner product is not scaled to make the matching scores more polarized.

Figure 7: Qualitative comparison of AdaFace with state-of-the-art face encoders. AdaFace generates images that maintain the highest authenticity of the subjects, while still follow the target prompts.

3 Experiments

3.1 Dataset and Training Details

We trained AdaFace on a combination of two face datasets: Flickr-Faces-HQ (FFHQ) [Karras et al., 2019], which comprises 70,000 images, and VGGFace2-HQ [Cao et al., 2018], which comprises 336,567 images after filtering. Face masks were generated using the BiSeNet face segmentation model [Yu et al., 2018]. The distilled Face2Image model is Ada2Face [Papantoniou et al., 2024], as it is able to generate authentic and diverse face images. The training employed the Prodigy optimizer [Mishchenko and Defazio, 2024] with d_coef=2 (akin to the learning rate in other optimizers) during face distillation, and d_coef=0.5 during composition distillation. Batch sizes were set to 4 and 3 for the two stages, respectively, with a gradient accumulation of 2. The model was trained with 240,000 iterations in the face distillation stage and 120,000 iterations in the composition distillation stage. During face distillation, the loss reached a plateau twice, resulting in two dynamic expansions of the model capacity. Eventually, the attention layers in the trained prompt inverter were expanded with multipliers of $(8x, 8x, 8x, 4x, 4x, ..., 4x)$ relative to the original CLIP text encoder. This resulted in a total of 2M parameters, in contrast to the 1.2M parameters of the original model.

In addition, we collected the images of 23 celebrities, each with 9 10 images, as the evaluated subjects. These celebrities include actors, singers and internet celebrities on Instagram. This dataset will be released along with the code.

3.2 Qualitative Comparisons

We compared AdaFace with a few state-of-the-art face encoders, including InstantID [Wang et al., 2024], ConsistentID [Huang et al., 2024] and PuLID [Guo et al., 2024b]. The input were images from our celebrity-23 dataset.

The results presented in Figure 7 demonstrate that AdaFace produces images that not only exhibit high authenticity of the subjects but also show good consistency with the text prompts. In comparison, other models often fall short in generating images that are either less authentic or less compositional. For instance, InstantID tends to produce overly stylized images with significant variability in authenticity across different subjects. PuLID, while generating aesthetically pleasing images, achieves slightly lower authenticity levels compared to AdaFace. Despite also utilizing a text-space approach,

Figure 8: Comparison of AdaFace with ID-Animator on personalized video generation. AdaFace generates videos with higher authenticity and compositionality.

ConsistentID has the least compositional output among the models evaluated, largely due to the absence of compositional training in its ID embeddings.

In addition, we plugged AdaFace into AnimateDiff, and generated personalized videos of celebrities under various compositional prompts. The results are shown in Figure 1. Figure 8 compares with a recent method ID-Animator [He et al., 2024]. AdaFace generated videos with high authenticity and compositionality, while ID-Animator usually produces videos with less authentic subjects.

3.3 Quantitative Evaluations

To assess the performance of AdaFace quantitatively, we evaluated a few baseline methods and AdaFace, on the “celebrity-23" images and DreamBench compositional prompts, comparing AdaFace with two baseline methods PuLID and InstantID. First, we measured the face similarity using the cosine similarity between the ArcFace embedding of the generated images and reference images. In addition, the CLIP-Text (CLIP-T) metric determines the consistency of the generated images with the prompts. The DINO and CLIP-I metrics are less indicative and are only for reference. The results, detailed in Table 1, show that AdaFace achieved comparable face similarity and prompt consistency scores to PuLID, and slightly outperformed InstantID. Note that the results of AdaFace is achieved on the original Stable Diffusion 1.5 model weight, which usually leads to much lower composition scores than other fine-tuned SD 1.5 model weights, such as RealisticVision.

	ArcFace (subj)	**CLIP-T (comp)**	DINO	CLIP-I
DB	0.349	**0.324**	0.470	0.656
TI	0.326	0.250	0.508	0.675
PuLID	0.468	**0.280**	0.512	0.630
InstantID	0.455	0.257	0.472	0.595
Ada	**0.476**	0.270	0.544	0.670
-Comp	0.505	0.235	0.598	0.685

Table 1: Quantitative evaluation on the “celebrity-23" images and DreamBench compositional prompts. **-Comp** is the model trained only with the face distillation stage.

As an ablation study, we list the performance of the AdaFace model without composition distillation. It can be seen that the face authenticity is slightly reduced after composition distillation, however, the generated images become much more consistent with the prompts.

4 Conclusions and Discussions

In this work, we present AdaFace, a versatile face encoder that maps human faces into the text prompt space. AdaFace is trained with a low compute budget and achieves high authenticity and compositionality in zero-shot generation of subject images. We demonstrate the effectiveness of AdaFace by showcasing the generated images and videos of celebrities under various compositional prompts. Additionally, our quantitative evaluations further underscore its performance.

A notable limitation of AdaFace is that the authenticity of the output face embeddings are constrained by the Face2Image model it distills from. However, this limitation can be addressed by distilling on more powerful Face2Image models and expanding the model capacity. For future work, we would extend the AdaFace method to object images. For instance, applying AdaFace distillation techniques to IP-Adapter [Ye et al., 2023] could enable the generation of both human and object images.

References

References

Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. VGGFace2: a dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74, Los Alamitos, CA, USA, May 2018. IEEE Computer Society. doi: 10.1109/FG.2018.00020. URL `https://doi.ieeecomputersociety.org/10.1109/FG.2018.00020`.

J. Chen, J. YU, C. GE, L. Yao, E. Xie, Z. Wang, J. Kwok, P. Luo, H. Lu, and Z. Li. PixArt-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *The twelfth international conference on learning representations*, 2024a. URL `https://openreview.net/forum?id=eAKmQPe3m1`.

W. Chen, J. Zhang, J. Wu, H. Wu, X. Xiao, and L. Lin. ID-Aligner: Enhancing Identity-Preserving Text-to-Image Generation with Reward Feedback Learning, Apr. 2024b. URL `http://arxiv.org/abs/2404.15449`. arXiv:2404.15449 [cs].

J. Deng, J. Guo, N. Xue, and S. Zafeiriou. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. pages 4690–4699, 2019. URL `https://openaccess.thecvf.com/content_CVPR_2019/html/Deng_ArcFace_Additive_Angular_Margin_Loss_for_Deep_Face_Recognition_CVPR_2019_paper.html`.

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, Aug. 2022a. URL `http://arxiv.org/abs/2208.01618`. arXiv:2208.01618 [cs].

R. Gal, O. Patashnik, H. Maron, A. H. Bermano, G. Chechik, and D. Cohen-Or. StyleGAN-NADA: CLIP-guided domain adaptation of image generators. *ACM Trans. Graph.*, 41(4), July 2022b. ISSN 0730-0301. doi: 10.1145/3528223.3530164. URL `https://doi.org/10.1145/3528223.3530164`. Number of pages: 13 Place: New York, NY, USA Publisher: Association for Computing Machinery tex.articleno: 141 tex.issue_date: July 2022.

Y. Guo, C. Yang, A. Rao, Z. Liang, Y. Wang, Y. Qiao, M. Agrawala, D. Lin, and B. Dai. AnimateDiff: Animate your personalized text-to-image diffusion models without specific tuning. In *The twelfth international conference on learning representations*, 2024a. URL `https://openreview.net/forum?id=Fx2SbBgcte`.

Z. Guo, Y. Wu, Z. Chen, L. Chen, and Q. He. PuLID: Pure and Lightning ID Customization via Contrastive Alignment, Apr. 2024b. URL `http://arxiv.org/abs/2404.16022`. arXiv:2404.16022 [cs].

Y. Han, J. Zhu, K. He, X. Chen, Y. Ge, W. Li, X. Li, J. Zhang, C. Wang, and Y. Liu. Face Adapter for Pre-Trained Diffusion Models with Fine-Grained ID and Attribute Control, May 2024. URL `http://arxiv.org/abs/2405.12970`. arXiv:2405.12970 [cs].

X. He, Q. Liu, S. Qian, X. Wang, T. Hu, K. Cao, K. Yan, and J. Zhang. ID-Animator: Zero-Shot Identity-Preserving Human Video Generation, May 2024. URL `http://arxiv.org/abs/2404.15275`. arXiv:2404.15275 [cs].

J. Ho, A. Jain, and P. Abbeel. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html`.

J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pages 7132–7141, 2018. doi: 10.1109/CVPR.2018.00745.

J. Huang, X. Dong, W. Song, H. Li, J. Zhou, Y. Cheng, S. Liao, L. Chen, Y. Yan, S. Liao, and X. Liang. ConsistentID: Portrait Generation with Multimodal Fine-Grained Identity Preserving, Apr. 2024. URL `http://arxiv.org/abs/2404.16771`. arXiv:2404.16771 [cs].

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 (CVPR)*, June 2019.

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 *Conference on computer vision and pattern recognition 2023*, 2023.

N. Kumari, B. Zhang, R. Zhang, E. Shechtman, and J.-Y. Zhu. Multi-Concept Customization of Text-to-Image Diffusion, Dec. 2022. URL `http://arxiv.org/abs/2212.04488`. arXiv:2212.04488 [cs].

K. Mishchenko and A. Defazio. Prodigy: An Expeditiously Adaptive Parameter-Free Learner, Mar. 2024. URL `http://arxiv.org/abs/2306.06101`. arXiv:2306.06101 [cs, math, stat].

A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models, Mar. 2022. URL `http://arxiv.org/abs/2112.10741`. arXiv:2112.10741 [cs].

F. P. Papantoniou, A. Lattas, S. Moschoglou, J. Deng, B. Kainz, and S. Zafeiriou. Arc2Face: A Foundation Model of Human Faces, Mar. 2024. URL `http://arxiv.org/abs/2403.11641`. arXiv:2403.11641 [cs].

W. Peebles and S. Xie. Scalable Diffusion Models with Transformers, Mar. 2023. URL `http://arxiv.org/abs/2212.09748`. arXiv:2212.09748 [cs].

D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *The twelfth international conference on learning representations*, 2024. URL `https://openreview.net/forum?id=di52zR8xgf`.

R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-Resolution Image Synthesis with Latent Diffusion Models, Apr. 2022. URL `http://arxiv.org/abs/2112.10752`. arXiv:2112.10752 [cs].

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, Aug. 2022. URL `http://arxiv.org/abs/2208.12242`. arXiv:2208.12242 [cs].

C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, R. Gontijo-Lopes, B. K. Ayan, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. Oct. 2022.

J. Shi, W. Xiong, Z. Lin, and H. J. Jung. InstantBooth: Personalized Text-to-Image Generation without Test-Time Finetuning, Apr. 2023. URL `http://arxiv.org/abs/2304.03411`. arXiv:2304.03411 [cs].

X. Sui, S. Li, X. Geng, Y. Wu, X. Xu, Y. Liu, R. Goh, and H. Zhu. CRAFT: Cross-attentional flow transformer for robust optical flow. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2022.

Z. Teed and J. Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, editors, *Computer vision – ECCV 2020*, pages 402–419, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58536-5.

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*, Siggraph '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701597. doi: 10.1145/3588432.3591506. URL `https://doi.org/10.1145/3588432.3591506`. Number of pages: 11 Place: , Los Angeles, CA, USA, tex.articleno: 12.

M. Toker, H. Orgad, M. Ventura, D. Arad, and Y. Belinkov. Diffusion Lens: Interpreting Text Encoders in Text-to-Image Pipelines, Mar. 2024. URL `http://arxiv.org/abs/2403.05846`. arXiv:2403.05846 [cs] version: 1.

Q. Wang, X. Bai, H. Wang, Z. Qin, A. Chen, H. Li, X. Tang, and Y. Hu. InstantID: Zero-shot Identity-Preserving Generation in Seconds, Feb. 2024. URL `http://arxiv.org/abs/2401.07519`. arXiv:2401.07519 [cs].

X. Wang, T. Chen, Q. Ge, H. Xia, R. Bao, R. Zheng, Q. Zhang, T. Gui, and X. Huang. Orthogonal subspace learning for language model continual learning. In *Findings of the association for computational linguistics: EMNLP 2023*, Singapore, Dec. 2023. Association for Computational Linguistics.

Y. Wei, Y. Zhang, Z. Ji, J. Bai, L. Zhang, and W. Zuo. ELITE: Encoding Visual Concepts into Textual Embeddings for Customized Text-to-Image Generation. In *ICCV 2023*. arXiv, Feb. 2023. doi: 10.48550/arXiv.2302.13848. URL `http://arxiv.org/abs/2302.13848`. arXiv:2302.13848 [cs].

H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang. IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models, Aug. 2023. URL `http://arxiv.org/abs/2308.06721`. arXiv:2308.06721 [cs].

C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang. BiSeNet: Bilateral segmentation network for real-time semantic segmentation. In V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, editors, *Computer vision – ECCV 2018*, pages 334–349, Cham, 2018. Springer International Publishing. ISBN 978-3-030-01261-8.

H. Zhang, J. Zhou, Y. Lu, M. Guo, P. Wang, L. Shen, and Q. Qu. The Emergence of Reproducibility and Consistency in Diffusion Models, Feb. 2024. URL `http://arxiv.org/abs/2310.05264`. arXiv:2310.05264 [cs].

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims in the abstract and introduction accurately reflect the paper's contributions and scope, as the detailed results, discussions, and conclusions align with and support the initial claims.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations of the work are discussed in the "Conclusions and Discussion" section.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: For each theoretical result, the paper provides a plenty of experimental support.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All the experimental details are clearly stated in the paper and the code will be made publicly available.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: All data and code will be made publicly available.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Please refer to the "Implementation Detail" section in the main paper.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: We evaluated on a diverse set of 30 celebrities, each with around 50 prompts, which is sufficient to reflect the model's performance.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We use 2 A6000 GPUs, each with 48G of memory.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conforms in every respect with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discussed the positive impacts, including its potential use in entertainment and art, video games and virtual reality. Additionally, its potential use for educational purposes in historical recreation, such as recreating faces of historical figures or enhancing documentaries, bringing history to life. We also pointed out potential negative impacts, including privacy violations. There is a risk of creating and using images of individuals without their consent. Moreover, misinformation and deepfakes are among the most concerning impacts, with the creation of deepfake videos that could be used to spread misinformation and manipulate public opinion. We also highlighted security concerns, as the technology could be used to bypass facial recognition systems for fraudulent purposes, posing significant security challenges. The authors will join in the effort for possible mitigation by providing gated release of models.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification: The work describes basic safeguards implemented for the responsible release of models, particularly focusing on preventing misuse. We have incorporated filters that specifically exclude NSFW (Not Safe for Work) keywords in the generation prompts, such as 'nude,' 'naked,' 'nsfw,' 'topless,' and 'bare breasts.' This approach helps mitigate the risk of generating inappropriate or sensitive content."

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: In our paper, we have ensured proper attribution for all assets used, such as code, data, and models, by citing the related papers and sources from which these assets were derived. Additionally, we have adhered to the licensing terms and conditions of each asset, as detailed in the respective citations.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: New assets introduced in the paper are well documented. The code is accompanied by usage documentation and is embedded with detailed comments to ensure clarity and ease of use for future researchers. Additionally, videos are provided alongside a description of the files and a list of prompts used for their generation, which enhances transparency and replicability of the results.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.