StableIdentity: Inserting Anybody into Anywhere at First Sight

Qinghe Wang[1*], Xu Jia[1*], Xiaomin Li[1], Taiqing Li[1], Liqian Ma[2], Yunzhi Zhuge[1], Huchuan Lu[1]
[1]**Dalian University of Technology, **[2]**ZMO AI Inc.**

https://qinghew.github.io/StableIdentity

Abstract

Recent advances in large pretrained text-to-image models have shown unprecedented capabilities for high-quality human-centric generation, however, customizing face identity is still an intractable problem. Existing methods cannot ensure stable identity preservation and flexible editability, even with several images for each subject during training. In this work, we propose StableIdentity, which allows identity-consistent recontextualization with just one face image. More specifically, we employ a face encoder with an identity prior to encode the input face, and then land the face representation into a space with an editable prior, which is constructed from celeb names. By incorporating identity prior and editability prior, the learned identity can be injected anywhere with various contexts. In addition, we design a masked two-phase diffusion loss to boost the pixel-level perception of the input face and maintain the diversity of generation. Extensive experiments demonstrate our method outperforms previous customization methods. In addition, the learned identity can be flexibly combined with the off-the-shelf modules such as ControlNet. Notably, to the best knowledge, we are the first to directly inject the identity learned from a single image into video/3D generation without finetuning. We believe that the proposed StableIdentity is an important step to unify image, video, and 3D customized generation models.

1 Introduction

With the boom in diffusion models [29, 28, 47], customized generation has garnered widespread attention [31, 45, 18]. This task aims to inject new subject (e.g., identity) into the text-to-image models and generate images with consistent subjects in various contexts while aligning the input text prompt. For example, users can upload their photos to obtain interesting pictures, such as "wearing a Superman outfit".

The success of customized generation can facilitate many applications such as personalized portrait photos [20], virtual try-on [5] and art & design [26].

However, existing customization methods solve this task by either finetuning the part/all parameters of the model or learning a generic encoder. Parameter finetuning methods [12, 8, 30] take a long time to search optimal parameters, but often return an inaccurate trivial solution for representing the identity. Especially if only with a single image, these methods tend to overfit the input, resulting in editability degradation. Alternatively, the encoder-based methods [43, 42] require large-scale datasets for training and struggle to capture distinctive identity and details. Moreover, the identity learned by current methods is susceptible to be inconsistent with the target identity in various contexts. Therefore, there is an urgent need to propose a new framework to address the enormous challenges (e.g., unstable identity preservation, poor editability, inefficiency) faced by this task.

Here we are particularly interested in customized generation for human under one-shot training setting, and how to store identity information into word embeddings, which can naturally integrate with text prompts. We believe prior knowledge can help for this task. On the one hand, face recognition task [36] has been fully explored and the identity-aware ability of pretrained models can be exploited. On the other hand, text-to-image models, which are trained on massive internet data, can generate images with celeb names in various contexts, thus these names contain rich editability prior. Using these priors can alleviate these challenges, and some methods [6, 45] have made preliminary attempts.

In this work, we propose StableIdentity which incorporates identity prior and editability prior into the human-centric customized generation. Specifically, an encoder pretrained on face recognition task is introduced to capture identity representation. Celeb names are collected to construct an embedding space as a prior identity distribution for customized generation. To encourage the target identity to perform like celeb names in pretrained diffusion model, we further land the identity representation into the prior space. Furthermore, to learn more stable identity and fine-grained reconstruction, we design a masked two-phase diffusion loss, which assigns specialized objectives in the early and late phases of denoising process respectively. Extensive experiments show StableIdentity performs favorably against state-of-the-art methods

---

* Corresponding authors.

Figure 1: Given a single input image, the proposed *StableIdentity* can generate diverse customized images in various contexts. Notably, we present that the learned identity can be combined with ControlNet [48] and even injected into video (ModelScopeT2V [37]) and 3D (Lucid-Dreamer [19]) generation.

and we further analyse our superiority over several baselines of the same kind. The proposed method also shows stable generalization ability, which can directly collaborate with the off-the-shelf image/video/3D models as shown in Figure 1.

Our contributions can be summarized as follows:

- We propose StableIdentity, which incorporates identity prior and editability prior to enable identity-consistent recontextualization with just one face image.

- We design a masked two-phase diffusion loss to perceive pixel-level details and learn more stable identity for diverse generation.

- Extensive experiments show that our method is effective and prominent. Remarkably, our method can not only combine with image-level modules, but also unlock the generalization ability that the identity learned from a single image can achieve identity-consistent customized video/3D generation without finetuning.

2 Related Work

2.1 Text-to-Image Diffusion Models

Diffusion models [15, 34] have exhibited overwhelming success in text-conditioned image generation, deriving numerous classical works [29, 24, 13]. Among them, Stable Diffusion [29] is widely used for its excellent open-source environment. In practice, Stable Diffusion can generate diverse

and exquisite images from Gaussian noises and text prompts with DDIM sampling [34]. Since the training dataset contains lots of celeb photos and corresponding names, Stable Diffusion can combine celeb names with different text prompts to generate diverse images. However, ordinary people cannot enjoy this "privilege" directly. Therefore, to democratize Stable Diffusion to broader users, many studies [6, 45, 4] have focused on the customized generation task.

2.2 Customized Generation

Currently, customized generation methods can be mainly divided into optimization-based and encoder-based methods. The former often require long time to optimize, while the latter need large-scale data and struggle to learn a distinctive identity. Given 3-5 images of the same subject, Textual Inversion [12] optimizes a new word embedding to represent the target subject. DreamBooth [30] finetunes the entire model to fit the target subject only. On the other hand, ELITE [38], InstantBooth [33] and IP-Adapter [43] introduce identity information into attention layers by learning an encoder. FastComposer [40] trains its encoder with the whole U-Net of Stable Diffsuion together to capture identities. There are also some methods that incorporate an encoder to assist the optimization-based methods [39], raising the performance ceiling. Celeb-Basis [45] collects 691 celeb names which are editable in Stable Diffusion to build a celeb basis by PCA [25]. The weight of basis is optimized based

Figure 2: Overview of the proposed *StableIdentity*. Given a single face image, we first employ a FR-ViT encoder and MLPs to capture identity representation, and then land it into our constructed celeb embedding space to better learn identity-consistent editability. In addition, we design a masked two-phase diffusion loss including $\mathcal{L}_{noise}$ and $\mathcal{L}_{rec}$ for training.

on the output of ArcFace encoder [7], a new identity's representation can be obtained by weighting the basis. However, the mentioned methods still perform imbalance on identity preservation and editability.

In comparison, our method exploits identity and editability prior to significantly ease the optimization process, and learns more stable identity with the proposed loss. Since Stable Diffusion is fixed, plug-and-play modules such as ControlNet [48] can be employed seamlessly. Furthermore, to the best knowledge, we are the first work to enable the learn identity from a single image injected into video [37] / 3D generation [19].

3 Method

Given a single face image, we aim to represent its identity via word embeddings as shown in Figure 2, to implement identity-consistent recontextualization under various text prompts. To achieve this, we incorporate identity prior and editability prior (See Sec. 3.2) and propose a masked two-phase diffusion loss (See Sec. 3.3).

3.1 Preliminary

In this work, we adopt the pretrained Stable Diffusion [29] as our text-to-image model (denoted as SD). SD consists of three components: a VAE ($\mathcal{E}$, $\mathcal{D}$) [11], a denoising U-Net ϵ_θ and a CLIP text encoder e_{text} [27]. Benefiting from the high-quality reconstruction of VAE, the diffusion process of input image x is performed in the latent space z ($z = \mathcal{E}(x)$). Specifically, at random timestep t ($t \in [1, 1000)$), z_t can be sampled as a weighted combination z_0 and a random noise ϵ ($\epsilon \sim \mathcal{N}(0, \mathbf{I})$):

$$z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, \tag{1}$$

where $\bar{\alpha}_t$ is a predefined hyperparameter set. Meanwhile, given text prompts p, the tokenizer of e_{text} divides and encodes p into l integer tokens sequentially. Then, the embedding layer in e_{text} obtains a corresponding embedding group $g = [v_1, ..., v_l], v_i \in \mathbb{R}^d$ which consists of l word embeddings by looking up the dictionary. After that, the text transformer τ_{text} of e_{text} further represents g to guide model to generate images conforming to the given text prompts p. With latent z_t, the training process is optimized by:

$$\mathcal{L}_{noise} = \mathbb{E}_{z,g,\epsilon,t} \left[\| \epsilon - \epsilon_\theta(z_t, t, \tau_{text}(g)) \|_2^2 \right] \tag{2}$$

3.2 Model Architecture

Identity Prior. Existing methods extract subject information commonly with CLIP image encoder, which is pretrained for learning high-level semantics, lacks detailed identity perception. Therefore, we employ a ViT [9] encoder finetuned for face recognition task (denote as FR-ViT) to reap ID-aware representation I from the input image.

To maintain the generalizability and editability of the vanilla SD, we fix the FR-ViT encoder and SD. Following [6, 45], we only project I into two word embeddings $[v_1', v_2']$ with MLPs:

$$[v_1', v_2'] = MLPs(I) \tag{3}$$

Benefiting the identity prior knowledge, we can inject facial features from the input image into diffusion model more efficiently without additional feature injection.

Editability Prior. Since SD is trained on large-scale internet data, using celeb names can generate images with prompt-consistent identity. Therefore, we posit that the celeb names constitute a space with editability prior. We consider 691 celeb names [45] as sampling points in this space and intend to represent this space distribution with the mean and standard deviation of their word embeddings. However, in practice, the tokenizer decomposes unfamiliar word into multiple

Figure 3: We present the predicted $\hat{z}_0$ from z_t at various timestep t. $\hat{z}_0$ at $t = \{100, 200\}$, similar to $t = 300$, are omitted for brevity.

tokens (e.g., *Deschanel* $\rightarrow$ [561, 31328, 832]), consequently the number of tokens produced by different celeb names may not be equal. To find an editable space with a uniform dimension, we select celeb names consisting only of first name and last name, and each word corresponds to only one token (e.g., *Tom Cruise* $\rightarrow$ [2435, 6764]). Eventually we obtain 326 celeb names and encode them into the corresponding word embeddings $C \in \mathbb{R}^{326 \times d}$.

To master the identity-consistent recontextualization ability like celeb embeddings, we employ AdaIN [10] to incorporate the editablity prior and land $[v'_1, v'_2]$ into celeb embedding space:

$$v_i^* = \sigma(C)(\frac{v'_i - \mu(v'_i)}{\sigma(v'_i)}) + \mu(C), for \ i = 1, 2 \quad (4)$$

where $\mu(v'_i), \sigma(v'_i)$ are scalars. $\mu(C) \in \mathbb{R}^d$, $\sigma(C) \in \mathbb{R}^d$ are vectors, since each dimension of C has a different distribution. With this editablity prior, the learned embeddings $[v_1^*, v_2^*]$ are closer to the celeb embedding space than baselines as shown in Figure 7, which improves editablity elegantly and effectively. In addition, it also constrains the optimization process within the celeb embedding space and prevents drifting towards other categories.

3.3 Model Training

Two-Phase Diffusion Loss. In addition to the architecture design, we rethink the training objective of diffusion models. The vanilla training loss $\mathcal{L}_{noise}$ excites the denoising U-Net ϵ_θ to predict the noise ϵ contained in the input z_t at any time t, and the introduced ϵ is randomly sampled each time. Therefore, such an objective function only implicitly and inefficiently learns the identity in the input image.

DDIM [34] proposes a denoised observation predicted by a variant of Eq. 1: $\hat{z}_0 = \frac{z_t - \sqrt{1 - \bar{\alpha}_t} \epsilon_\theta}{\sqrt{\bar{\alpha}_t}}$. A naive idea is to replace $\mathcal{L}_{noise}$ with the mean squared error between the predicted $\hat{z}_0$ and the real z_0 [39]: $\mathcal{L}_{rec} = \mathbb{E}_{z,g,\epsilon,t} \left[\|\hat{z}_0 - z_0\|_2^2\right]$, which can explicitly optimize the reconstruction for z_0. However, we observe that as timestep increases, predicted $\hat{z}_0$ becomes more difficult to approximate the true distribution of z_0 as shown in Figure 3. Therefore, for larger timestep, $\mathcal{L}_{rec}$ becomes less meaningful and even misleads the model to focus excessively on pixel-level reconstruction. To this end, we

propose two-phase diffusion loss divided by timestep αT:

$$\mathcal{L}_{diffusion} = \begin{cases} \mathbb{E}_{z,g,\epsilon,t} \left[\|\epsilon - \epsilon_\theta(z_t, t, \tau_{text}(g))\|_2^2\right] & t \geq \alpha T, \\ \mathbb{E}_{z,g,\epsilon,t} \left[\|\hat{z}_0 - z_0\|_2^2\right] & t < \alpha T. \end{cases} \quad (5)$$

Empirically, the division parameter $\alpha \in [0.4, 0.6]$ yields good results that balance identity preservation and diversity ($\alpha = 0.6$ as default). Using $\mathcal{L}_{noise}$ at the early phase of denoising process that decides the layout of generated image [21, 40, 23] can allow the learned identity to adapt to diverse layouts, while using $\mathcal{L}_{rec}$ at the late phase can boost the pixel-level perception for input image to learn more stable identity.

Masked Diffusion Loss. To prevent learning irrelevant background, we also employ the masked diffusion loss [1, 39]. Specifically, we use a pretrained face parsing model [44] to obtain the face mask M_f and hair mask M_h of the input image. The training loss is calculated in the face area and hair area respectively:

$$\mathcal{L} = M_f \odot \mathcal{L}_{diffusion} + \beta M_h \odot \mathcal{L}_{diffusion}. \quad (6)$$

In our experiments, we set $\beta = 0.1$ as default.

4 Experiments

4.1 Experimental Setting

Implementation Details. Our experiments are based on Stable Diffusion 2.1-base. The FR-ViT is a ViT-B/16 encoder finetuned for face recognition task. For an input single image, we use color jitter, random shift, random crop and random resize as data augmentations. The learning rate and batch size are set to $5e - 5$ and 1. The MLPs are trained for 450 steps (4 mins). The placeholders $v_1^* \ v_2^*$ of prompts such as "v_1^* v_2^* wearing a spacesuit", "latte art of $v_1^* \ v_2^*$" are omitted for brevity in this paper. The scale of classifier-free guidance [14] is set to 8.5 by default. Our experiments are conducted on a single A800 GPU.

Dataset. We randomly select 70 non-celeb images from the FFHQ [16] and resize to 512×512 as our test dataset. To perform a comprehensive evaluation, we employ 40 test prompts which cover actions, decorations, attributes, expressions and backgrounds [18].

Baselines. We compare the proposed method with baselines including the optimization-based methods: Textual Inversion [12], DreamBooth [30], Celeb-Basis [45] and the encoder-based methods: ELITE [38], FastComposer [40], IP-Adapter [43]. We prioritize using the official model released by each method. For Textual Inversion and DreamBooth, we use their Stable Diffusion 2.1 versions for a fair comparison.

Evaluation Metrics. Following DreamBooth [30], we calculate CLIP [27] visual similarity (**CLIP-I**) to evaluate high-level semantic alignment and text-image similarity (**CLIP-T**) to measure editablity. Besides, we calculate the **Face Similarity** by ArcFace [7] and **Face Diversity** [18, 39] by LPIPS [46] on detected face regions between the generated images and real images of the same ID. However, some baselines may generate completely inconsistent faces under various text prompts, which incorrectly raise face diversity. Therefore, we propose the **Trusted Face Diversity** by the

Figure 4: We present the qualitative comparisons with six baselines for different identities (including various races) and diverse text prompts (covering decoration, action, attribute, background, style). Our method achieves high-quality generation with consistent identity and outstanding editability (*Zoom-in for the best view*). We provide more results in supplementary material.

product of cosine distances from face similarity and face diversity for each pair images, to evaluate whether the generated faces are both diverse and similar. To measure the quality of generation, we randomly sample 70 celeb names to generate images with test prompts as pseudo ground truths and calculate Fréchet Inception Distance (**FID**) [22] between the generated images by the competing methods and pseudo ground truths.

4.2 Comparison

Qualitative Comparison. As shown in Figure 4, given a single image as input, we show the generation results with various prompts. Textual Inversion is optimized only with $\mathcal{L}_{noise}$, which leads to a trivial solution for identity in different contexts. DreamBooth finetunes the whole SD model to fit the input face, but still fails to learn similar identities (row $1_{th}, 5_{th}$) and tends to replicate the foreground face (row $2_{th}, 3_{th}$). The encoder-based methods ELITE and IP-Adapter only learn rough shape and attributes of the in-

put face, perform mediocrely in both identity preservation and editability. FastComposer finetunes a CLIP image encoder and the whole SD for learning identities, but suffers from low quality and artifacts (row $4_{th}, 5_{th}, 6_{th}$). Celeb-Basis also fails to learn accurate identity for recontextualization (row $1_{th}, 3_{th}$). Notably, when using "latte art of" as text prompt, all baselines either produce inconsistent identity or fail to get the desired style in row 6_{th}. In comparison, benefiting from the effectiveness of the proposed method, our results shows superiority both in terms of identity preservation and editablity.

Quantitative Comparison. In addition, we also report the quantitative comparison in Table 1. Some baselines like ELITE and IP-Adapter learn only facial structure and attributes, and are susceptible to generate frontal view, resulting in better CLIP-I. This metric focuses on high-level semantic alignment and ignores identity consistency. Therefore, these methods obtain worse face similarity (-24.54, -15.39 than ours) and trusted face diversity (-9.66, -3.95 than ours).

Table 1: Quantitative comparisons with baselines. ↑ indicates higher is better, while ↓ indicates that lower is better. The best result is shown in **<u>bold</u>**. Our method obtains the best results over the text consistency (i.e., CLIP-T), identity preservation (i.e., Face Similarity), diversity of generated faces (i.e., Trusted Face Diversity), and generation quality (i.e., FID).

	CLIP-I↑(%)	CLIP-T↑(%)	Face Sim.↑(%)	Face Div.↑(%)	Trusted Div.↑(%)	FID↓
Textual Inversion	61.30	28.23	31.30	37.03	10.75	28.64
DreamBooth	67.01	28.91	35.80	36.79	5.89	48.55
ELITE	73.94	26.43	12.58	25.55	5.35	84.32
FastComposer	72.32	28.87	36.91	28.84	13.90	47.98
IP-Adapter	**<u>85.14</u>**	23.67	21.73	25.61	11.06	78.95
Celeb-Basis	63.69	27.84	25.55	**<u>37.85</u>**	13.41	33.72
StableIdentity (Ours)	65.91	**<u>29.03</u>**	**<u>37.12</u>**	35.46	**<u>15.01</u>**	**<u>24.92</u>**

Table 2: Ablation study. We also present results with various division parameter α in the supplementary material.

	CLIP-T↑	Face Sim.↑	Trusted Div.↑	FID↓
CLIP Enc.	28.03	35.73	14.81	25.66
w/o AdaIN	24.81	**<u>47.81</u>**	13.73	48.73
w/o Mask	28.15	34.98	14.47	25.12
Only $\mathcal{L}_{noise}$	28.81	36.55	14.97	25.76
Only $\mathcal{L}_{rec}$	27.35	30.69	13.89	40.54
Ours	**<u>29.03</u>**	37.12	**<u>15.01</u>**	**<u>24.92</u>**

We also observe that the optimization-based methods Textual Inversion and DreamBooth fail to learn stable identities for recontextualization and tend to overfit to the input face, leading to poor trusted face diversity (-4.26, -9.12 than ours). Our method achieves best performance on vision-language alignment (CLIP-T), identity preservation (Face Sim.), identity-consistent diversity (Trusted Div.) and image quality (FID).

4.3 Ablation Study

We conduct a comprehensive ablation study across various settings as shown in Table 2 and Figure 5, 6. We employ the CLIP Image Encoder as a baseline which is commonly adopted in encoder-based methods. Following [33, 40], we use the CLS feature of CLIP encoder's last layer to extract identity information. In col 2 of Figure 5, it can be observed that the CLIP image encoder is mediocre for identity preservation (-1.39 than ours on Face Sim.). On the other hand, the setting of "w/o AdaIN" cannot explicitly learn editability and fails to limit the value range of the learned word embeddings. It tends to generate the frontal faces and fails to align the desired text prompt (col 3 in Figure 5), therefore it obtains high face similarity but poor CLIP-T, Trusted Div., and FID (-4.22, -1.28, -23.81 than ours).

Furthermore, we show the ablation results for the training loss. The masked diffusion loss has been proven effective [1, 39] and it does help focus foreground and prevent background leakage. The reconstruction of the "Only $\mathcal{L}_{noise}$" setting is inferior than ours and is prone to undesired changes and artifacts (col 3 in Figure 6), resulting lower identity preservation and image quality (i.e., -0.60, -0.84 than ours on Face Sim., FID). Due to the meaningless $\mathcal{L}_{rec}$ in the early phase of denoise process, the "Only $\mathcal{L}_{rec}$" setting only learns

Figure 5: Ablation study for model architecture. We show the results of using the CLIP image encoder and removing the AdaIN.

Figure 6: Ablation study for training loss. We present the visualization results of various loss settings.

mediocre identities with artifacts (col 4 in Figure 6) and leads to unsatisfactory face similarity, trusted diversity, and FID (-6.43, -1.12, -15.62 than ours). In comparison, the proposed masked two-phase diffusion loss shows best results, and the discussion of the division parameter α can be found in supplementary material.

5 Discussion

5.1 Downstream Applications

Pose-controlled Customized Image Generation. Since the pretrained Stable Diffusion is fixed, SD-based plug-and-play modules can collaborate with our method. ControlNet controls the pretrained SD to support additional input conditions such as keypoints, edge maps, etc. In this paper, we obtain pose images with human skeletons as condition by Open-Pose [3], as an example. As shown in the row 2 of Figure 1,

Figure 7: 2-D visualization of word embeddings using *t*-SNE with Celeb names, Textual Inversion, Celeb-Basis and our method.

Table 3: Comparison with baselines optimized in the word embedding space on training time, maximum and minimum values of learned embeddings.

	Training time	Max	Min
Celeb names	–	0.0551	-0.0558
Textual Inversion	43mins	0.7606	-0.9043
Celeb-Basis	8mins	0.1592	-0.1499
StableIdentity (Ours)	4mins	0.0557	-0.0520

Figure 8: Comparison of 3D generation based on LucidDreamer. We show the result of a celeb name "Tom Cruise" (prompt) as a standard and the results with the embeddings $[v_1^*, v_2^*]$ learned from competing methods (*Zoom-in for the best view*).

we demonstrate the integration of StableIdentity and Control-Net (SD2.1 version) which achieves simultaneous structure-controlled and identity-driven generation.

Zero-shot Identity-driven Video/3D Generation. Our method can be considered as introducing new identity for the dictionary of CLIP text encoder. Therefore, we believe that ideally, as long as the SD-based video and 3D generation models do not finetune the CLIP text encoder, the learned identity can be directly injected into these models.

ModelScopeT2V [37] is a text-to-video generation model which brings some temporal structures into the U-Net of SD2.1 and finetunes the U-Net on large-scale datasets [32, 2, 41]. We attempt to insert the learned identity into the unchanged CLIP text encoder without finetuning as shown in the row 3 of Figure 1. The generated video shows promising identity preservation and text alignment.

LucidDreamer [19] is a text-to-3D generation pipeline based on 3D Gaussian Splatting [17] and allows to sample directly with the pre-trained SD2.1, like us. Therefore, it can naturally collaborate with our method. In a similar way, we insert the learned identity into this pipeline, as shown in the row 4 of Figure 1. The generated results achieve stable identity, high fidelity and geometry consistency. The result of "wearing a golden crown" exhibits precise geometric structures and realistic colors and the "as oil painting" obtains the desired style, a 3D portrait oil painting that does not exist in reality.

Overall, our method can effortlessly enable prompt-consistent identity-driven video/3D generation with the off-the-shelf text-to-video/text-to-3D models. We show more results of video/3D in the supplementary material.

5.2 Word-Embedding-Based Methods Analysis

Considering that Textual Inversion, Celeb-Basis and our method are all optimized in the word embedding space, we further analyze 70 embeddings learned by these methods from different perspectives. To match the dimension of word embeddings, Textual Inversion is conducted with 2-word version and Celeb-Basis is implemented with SD2.1 for analysis.

To intuitively show the difference between the distributions of learned embeddings and celeb embeddings, we use the *t*-SNE [35] to visualize word embeddings in Figure 7. "Celeb names" denotes the word embeddings corresponding to the collected 326 celeb names. It can be observed that the distribution of ours is more compact with fewer outliers and closer to the real distribution of celeb names, achieving the best identity-consistent editability. Besides, we compare the max & min values of the learned embeddings and training time in Table 3. Our method is faster than existing methods of the same kind, and the value range is closest to real celeb embeddings.

Furthermore, to examine the generalization ability of these methods, we present 3D generation results with the learned identity embeddings directly using the mentioned 3D generation pipeline LucidDreamer in Figure 8. And we show a standard result using a celeb name "Tom Cruise" as a prompt. Obviously, our method achieves celeb-like results in every 3D view, which further demonstrates stable and strong generalization ability of our learned identity.

6 Conclusion

In this paper, we propose *StableIdentity*, a customized generation framework which can inject anybody into anywhere. The model architecture that integrates identity and editability prior allows the learned identity to master identity-consistent recontextualization ability. Besides, the designed masked two-phase diffusion loss enables the learned identity more stable. Extensive quantitative and qualitative experiments demonstrate the superiority of the proposed method. Surprisingly, our method can directly work with the plug-and-play SD-based modules such as ControlNet, and even can insert the learned identity into off-the-shelf video/3D generated models without finetuning to produce outstanding effects. We hope that our work can contribute to the unification of customization over image, video, and 3D generation tasks.

References

[1] Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. Break-a-scene: Extracting multiple concepts from a single image. *arXiv preprint arXiv:2305.16311*, 2023. 4, 6

[2] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738, 2021. 7

[3] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 6

[4] Li Chen, Mengyi Zhao, Yiheng Liu, Mingxu Ding, Yangyang Song, Shizun Wang, Xu Wang, Hao Yang, Jing Liu, Kang Du, et al. Photoverse: Tuning-free image customization with text-to-image diffusion models. *arXiv preprint arXiv:2309.05793*, 2023. 2

[5] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. *arXiv preprint arXiv:2307.09481*, 2023. 1

[6] Zhuowei Chen, Shancheng Fang, Wei Liu, Qian He, Mengqi Huang, Yongdong Zhang, and Zhendong Mao. Dreamidentity: Improved editability for efficient face-identity preserved image generation. *arXiv preprint arXiv:2307.00300*, 2023. 1, 2, 3

[7] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 3, 4

[8] Ziyi Dong, Pengxu Wei, and Liang Lin. Dreamartist: Towards controllable one-shot text-to-image generation via contrastive prompt-tuning. *arXiv preprint arXiv:2211.11337*, 2022. 1

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[10] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. *arXiv preprint arXiv:1610.07629*, 2016. 4

[11] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 3

[12] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 1, 2, 4

[13] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 2

[14] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 4

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2

[16] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 4

[17] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023. 7

[18] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. *arXiv preprint arXiv:2312.04461*, 2023. 1, 4

[19] Yixun Liang, Xin Yang, Jiantao Lin, Haodong Li, Xiaogang Xu, and Yingcong Chen. Luciddreamer: Towards high-fidelity text-to-3d generation via interval score matching. *arXiv preprint arXiv:2311.11284*, 2023. 2, 3, 7, 10

[20] Yang Liu, Cheng Yu, Lei Shang, Ziheng Wu, Xingjun Wang, Yuze Zhao, Lin Zhu, Chen Cheng, Weitao Chen, Chao Xu, et al. Facechain: A playground for identity-preserving portrait generation. *arXiv preprint arXiv:2308.14256*, 2023. 1

[21] Zhiheng Liu, Yifei Zhang, Yujun Shen, Kecheng Zheng, Kai Zhu, Ruili Feng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Cones 2: Customizable image synthesis with multiple subjects. *arXiv preprint arXiv:2305.19327*, 2023. 4

[22] Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. Are gans created equal? a large-scale study. *arXiv preprint arXiv:1711.10337*, 2017. 5

[23] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 4

[24] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2

[25] Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh,*

and Dublin philosophical magazine and journal of science, 2(11):559–572, 1901. 2

[26] Joern Ploennigs and Markus Berger. Ai art in architecture. *AI in Civil Engineering*, 2(1):8, 2023. 1

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3, 4

[28] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022. 1

[29] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1, 2, 3, 10

[30] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir 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*, pages 22500–22510, 2023. 1, 2, 4

[31] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Wei Wei, Tingbo Hou, Yael Pritch, Neal Wadhwa, Michael Rubinstein, and Kfir Aberman. Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. *arXiv preprint arXiv:2307.06949*, 2023. 1

[32] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 7

[33] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. *arXiv preprint arXiv:2304.03411*, 2023. 2, 6

[34] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2, 4

[35] Laurens Van Der Maaten. Accelerating t-sne using tree-based algorithms. *The journal of machine learning research*, 15(1):3221–3245, 2014. 7

[36] Mei Wang and Weihong Deng. Deep face recognition: A survey. *Neurocomputing*, 429:215–244, 2021. 1

[37] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023. 2, 3, 7, 10, 12

[38] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. *arXiv preprint arXiv:2302.13848*, 2023. 2, 4

[39] Zijie Wu, Chaohui Yu, Zhen Zhu, Fan Wang, and Xiang Bai. Singleinsert: Inserting new concepts from a single image into text-to-image models for flexible editing. *arXiv preprint arXiv:2310.08094*, 2023. 2, 4, 6

[40] Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *arXiv preprint arXiv:2305.10431*, 2023. 2, 4, 6

[41] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016. 7

[42] Yuxuan Yan, Chi Zhang, Rui Wang, Yichao Zhou, Gege Zhang, Pei Cheng, Gang Yu, and Bin Fu. Facestudio: Put your face everywhere in seconds. *arXiv preprint arXiv:2312.02663*, 2023. 1

[43] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 1, 2, 4

[44] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 325–341, 2018. 4

[45] Ge Yuan, Xiaodong Cun, Yong Zhang, Maomao Li, Chenyang Qi, Xintao Wang, Ying Shan, and Huicheng Zheng. Inserting anybody in diffusion models via celeb basis. *arXiv preprint arXiv:2306.00926*, 2023. 1, 2, 3, 4, 10

[46] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 4

[47] Chenshuang Zhang, Chaoning Zhang, Mengchun Zhang, and In So Kweon. Text-to-image diffusion model in generative ai: A survey. *arXiv preprint arXiv:2303.07909*, 2023. 1

[48] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 2, 3, 10

[49] Yiqun Zhang, Zhenyue Qin, Yang Liu, and Dylan Campbell. Detecting and restoring non-standard hands in stable diffusion generated images. *arXiv preprint arXiv:2312.04236*, 2023. 12

| Input Image | coding in front of a computer | wearing a Superman outfit | wearing a purple wizard outfit | playing the violin | with red hair |

Figure 9: More generated results with the proposed *StableIdentity* for different identities (including various races) under various contexts (covering decoration, action, attribute).

A More Visual Results

More Customized Results. As shown in Figure 9, it can be observed that *StableIdentity* can handle different races in various contexts. On the other hand, we also show the customized results with diverse artistic styles in Figure 10. Overall, the generated results have satisfactory identity preservation and editability, which demonstrates the effectiveness of the proposed method.

StableIdentity & Image/Video/3D Models. In addition, as shown in Figure 12, 13, we show more image/video/3D customized generation results with ControlNet [48], ModelScopeT2V [37], and LucidDreamer [19]. As introduced in Sec. 4.1, StableIdentity can be considered as introducing new identity for the dictionary of CLIP text encoder. Therefore, the learned identity can be naturally inserted into various contexts or even into video/3D generated models for identity-driven customized generation. Due to the limited performance of 3D generation, following [19], we only generate and edit in the head region, which can clearly demonstrates

whether the learned identity is successfully inserted or not. Impressive experimental results show that our method can be stably injected into image/video/3D generative models to achieve identity-consistent recontextualization. Furthermore, we also show more customized generation results using celeb photos as input as shown in 14.

More Comparisons. As shown in Figure 15, we further compare with baselines on decoration, action, background, style. Obviously, we achieve the best results for identity preservation, editability and image quality. DreamBooth, which seems to perform well (row 1,3,4), either overfits to the input image or fails to produce results similar to the target identity.

B Implementation Details

Filtering Celeb Names. As mentioned in Sec. 2, Celeb-Basis [45] collects 691 celeb names which are editable in Stable Diffusion [29]. We only filter out names consisting of only two words and then count the number of the corresponding tokens. The percentages of 2 words→{2,3,4,5,6}

Figure 10: Additional customized results with *StableIdentity* for diverse artistic styles.

| Input Image | a sand sculpture of | in Ghibli animation style | a pencil sketch of | a Ukiyo-e painting of | Fauvism painting of |

Figure 11: Parameter analysis for the division parameter α.

tokens are 56%, 27%, 13%, 3%, and 0.3% respectively. To obtain a more accurate prior space, we choose the setting of 2 words→2 tokens, which has more sampling points.

Division Parameter α. As shown in Figure 11, we present the effect of different α in $[0, 0.1, \cdots, 1]$. Empirically, $\alpha \in [0.4, 0.6]$ shows better identity preservation, editability and image quality. When α is larger, meaningless reconstruction will lead to performance degradation.

C Limitations

Although the proposed method achieves outstanding performance for customization generation of new identities and can collaborate with the off-the-shelf image/video/3D models, it

Figure 12: Pose-controlled customized image generation (StableIdentity & ControlNet) and zero-shot identity-driven customized video generation (StableIdentity & ModelScopeT2V).

still faces some limitations. (1) Since we only act in the word embedding space and fix the Stable Diffusion (SD), we inherit not only the excellent performance of SD, but also some drawbacks, such as hand anomalies [49]. (2) Existing text-to-video generation models can generate with diverse contexts, but is still immature for human-centric generation [37]. It leads to limited performance for the video customization generation.

Figure 13: Zero-shot identity-driven customized 3D generation (StableIdentity & LucidDreamer). As mentioned in Sec. 4.1, we omit the placeholders v_1^* v_2^* of prompts such as "v_1^* v_2^* wearing glasses" for brevity. Here, we use "v_1^* v_2^*" as the input prompt to show the 3D reconstruction for the learned identities.

Input Image   latte art of   & ControlNet: holding a lollipop

& ModelScopeT2V: with sunglasses driving   & LucidDreamer: wearing a hat

Input Image   wearing Ironman suit   & ControlNet: wearing Captain America costume

& ModelScopeT2V: wearing headphones and eyeglasses   & LucidDreamer: v_1^* v_2^*

Input Image   wearing spiderman suit   & ControlNet: in front of Eiffel Tower

& ModelScopeT2V: drinking water   & LucidDreamer: made out of toy bricks

Input Image   playing the guitar   & ControlNet: cooking in the kitchen

& ModelScopeT2V: wiping face   & LucidDreamer: in comic style

Figure 14: More image/video/3D customized generation results for celeb photos as input.

Figure 15: More qualitative comparisons for different identities (including various races) with diverse text prompts (covering decoration, action, background, style). Our method shows best performance for identity preservation and editability (*Zoom-in for the best view*).