# RefDrop: Controllable Consistency in Image or Video Generation via Reference Feature Guidance

**Jiaojiao Fan**
Georgia Tech
sbyebss@gmail.com

**Haotian Xue**
Georgia Tech
htxue.ai@gatech.edu

**Qinsheng Zhang**
NVIDIA
qsh.zh27@gmail.com

**Yongxin Chen**
Georgia Tech
yongchen@gatech.edu

## Abstract

There is a rapidly growing interest in controlling consistency across multiple generated images using diffusion models. Among various methods, recent works have found that simply manipulating attention modules by concatenating features from multiple reference images provides an efficient approach to enhancing consistency without fine-tuning. Despite its popularity and success, few studies have elucidated the underlying mechanisms that contribute to its effectiveness. In this work, we reveal that the popular approach is a linear interpolation of image self-attention and cross-attention between synthesized content and reference features, with a constant rank-1 coefficient. Motivated by this observation, we find that a rank-1 coefficient is not necessary and simplifies the controllable generation mechanism. The resulting algorithm, which we coin as `RefDrop`, allows users to control the influence of reference context in a direct and precise manner. Besides further enhancing consistency in single-subject image generation, our method also enables more interesting applications, such as the consistent generation of multiple subjects, suppressing specific features to encourage more diverse content, and high-quality personalized video generation by boosting temporal consistency. Even compared with state-of-the-art image-prompt-based generators, such as IP-Adapter, `RefDrop` is competitive in terms of controllability and quality while avoiding the need to train a separate image encoder for feature injection from reference images, making it a versatile plug-and-play solution for any image or video diffusion model. Our project webpage is https://sbyebss.github.io/refdrop/.

## 1 Introduction

Large-scale diffusion models have demonstrated remarkable capabilities in aiding content creation for artists [43, 7, 3]. Numerous text-to-image models are expediting content production in various domains, including advertising and art studios. Similarly, video generation models have shown significant advancements recently [17, 9, 18, 24, 52, 6, 4, 23]. However, enhancing these models to better support artistic creativity requires improved controllability, particularly in content consistency. This paper explores consistency from two perspectives: 1) controlling subject consistency across multiple images, and 2) maintaining subject consistency across multiple frames within a video.

We name a few tasks where the controllable consistency is crucial in AI content generation. In image generation for storytelling [45, 39, 41, 27] or advertising, content creators often strive to produce consistent characters, a task that proves challenging with foundational generative models [55]. Personalization approaches based on fine-tuning [50] require a minimum of 5 to 10 images to achieve

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


https://doi.org/10.52202/079017-1058


Figure 1: RefDrop achieves controllable consistency in visual content synthesis for free. RefDrop exihibits great flexibility in (Upper) multi-subject consistency generation given one reference image, (Middle) blending different characters from multiple images seamlessly, (Buttom) enhancing temporal consistency for personalized video generation. RefDrop is short for "reference drop". We named our method RefDrop to metaphorically represent the process by which a drop of colored water influences a larger body of clear water.

satisfactory quality, and encoder-based methods [61, 59, 40] demand weeks of training with millions of images for a single diffusion model and lack transferability to other foundational models. On the other hand, diverse image generation is less addressed but persistently challenging. In this scenario, it is desired to *decrease* the consistency among image generations. For example, artists can sometimes seek to enhance diversity and avoid clichés, such as the stereotypical depiction of Barbie girls with curly blonde hair. For video generation, another challenging task is maintaining temporal consistency in video generation, yet most existing solutions are confined to video editing tasks [31], demanding high-quality input videos.

These emerging tasks motivate us to develop RefDrop, a **training-free**, **plug-and-play** method designed to provide flexible control over the consistency in image and video generation. Specifically, we modify the self-attention mechanism in the diffusion model UNet [49] architecture and introduce a coefficient to modulate the influence of a reference image on the generation process. Our contributions are outlined as follows:

1. We conduct a detailed analysis of popular consistency generation methods based on concatenated attention, revealing that their consistency is actually contributed by extra guidance applied implicitly.
2. Inspired by this finding, we propose Reference Feature Guidance (RFG), a natural extension that

explicitly controls the guidance from reference context in a precise and direct manner. Building upon RFG, we introduce RefDrop, a flexible and efficient approach to controlling consistency without the need for network fine-tuning or optimization. 3. Besides improvements in character consistency using a single reference image, RefDrop enables more creative applications with controllable consistency, including (i) seamless integration of distinct features into a single cohesive image (ii) suppressing specific features by negatively decreasing the consistency influenced by the reference context, thereby enhancing diversity in layout, accessories, and image style; (iii) high-quality personalized video generation by boosting temporal consistency, and minimizing facial distortions. 4. We conduct comprehensive experiments and demonstrate that RefDrop achieves a good balance between flexibility and effectiveness while being lightweight compared to existing works.

## 2 Related work

Among the works most similar to ours are IP-Adapter [67] and concatenated attention [62]. Our approach is closely related to IP-Adapter, as both methods utilize the sum of two decoupled attention outputs. However, while IP-Adapter modifies cross-attention and requires separate training of an image encoder to embed the reference image, we integrate the reference image directly into the self-attention layer without needing additional training. Furthermore, our reference images are *generated* by the same model, in contrast to IP-Adapter's reliance on externally sourced image. Both techniques permit the use of negative or positive coefficients for the reference image, but IP-Adapter may compromise text alignment [55] due to its reference image being intertwined with the text prompt during cross-attention. Additionally, the IP-Adapter requires separate training for different versions of the diffusion model, such as SD2.1 and SDXL. In contrast, RefDrop is a simple plug-and-play.

Concatenated attention, first introduced in video generation literature by Wu et al. [62] as spatio-temporal attention, injects temporal information into a T2I model. It has since been widely adopted for feature injection across various applications [37, 8, 25, 55] in content generation and video editing. This concept has evolved into Cross-Frame Attention [29], another prevalent technique used to inflate T2I models [69] for video generation. We will demonstrate later that our framework can replicate these two types of attention as special cases.

A concurrent work by Avrahami et al. [2] introduces a method called soft blending, which is quite similar to our RFG (5), but applies it to a different application: object dragging.

**Consistency in Image Generation** ConsiStory [55] and StoryDiffusion [71] are closely related to our work. They are training-free methods that employs concatenated attention to enhance consistency in generation. Our RFG framework is *orthogonal* to the techniques other than concatenated attention in those works, such as subject masking and attention dropout. Avrahami et al. [1] explores a fine-tuning-based method aimed at recovering tightly clustered images. Other approaches, such as those by [27, 12, 35], predominantly utilize a personalization process [50, 14, 32, 54] requiring multiple input images for training. Finally, several encoder-based methods [61, 59, 33, 51, 64, 30, 34] do not require additional training for new subjects. However, these methods necessitate days or weeks of initial training for the encoder and face limitations in adaptability to different versions of foundational generative models.

**Temporal-consistency in video generation** Concatenated attention [62, 47] and Cross-Frame Attention [29, 10] are popular techniques used to inflate T2I models for video generation. Wu et al. [63], Ren et al. [47] mitigate video flickering by applying a low-pass filter to noisy latent images, effectively removing disruptive high-frequency content. Many other methods are tailored for video editing tasks, and they either extract features from high-quality input videos to enhance the current generation [31, 70, 66, 65] or use them as references during editing [16, 11]. RefDrop improves temporal consistency directly within video generation, obviating the need for an input video.

## 3 Method

We first introduce how existing works achieve consistency generation by leveraging the concatenation of reference features in the self-attention block. Then we reformulate the concatenation as a linear interpolation of self-attention on synthesized content and cross-attention between generated and reference content with a constant rank-1 coefficient. We highlight that this specific coefficient is not a

 https://doi.org/10.52202/079017-1058

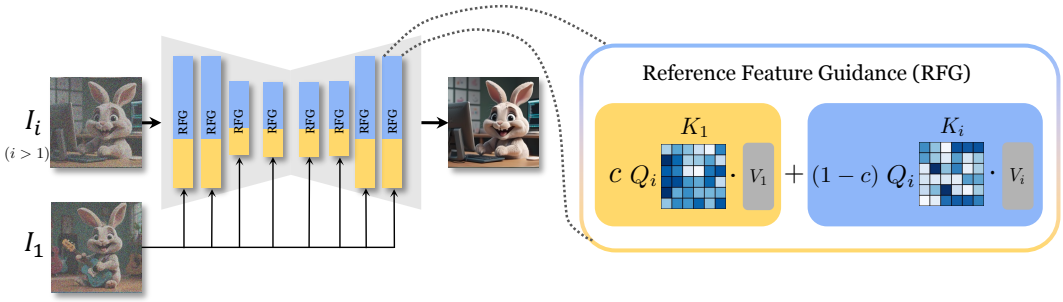


Figure 2: During each diffusion denoising step, we facilitate the injection of features from a *generated* reference image $I_1$ into the generation process of other images through `RFG`. The `RFG` layer produces a linear combination of the attention outputs from both the standard and referenced routes. A negative coefficient $c$ encourages divergence of $I_i$ from $I_1$, while a positive coefficient fosters consistency.

necessity, while linear interpolation is critical to minimizing the training-inference gap. Building upon these observations, we propose Reference Feature Guidance (`RFG`), an extension of concatenation attention that allows for flexible feature interpolation and extrapolation in attention modules. Based on `RFG`, we introduce `RefDrop`, a versatile method for controllable consistency generation across various applications.

### 3.1 Background

Self-attention in diffusion model networks operates by applying the attention mechanism [57] on synthesized latent features. A self-attention layer processes latent representations $X$ by passing them through linear projection layers to produce queries $Q = XW_Q$, keys $K = XW_K$, and values $V = XW_V$, which then undergo the attention operation as follows:

$$X' = \text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V, \tag{1}$$

where $X'$ is the output of self-attention operation, and $d$ is the feature dimension of projection matrices $W_Q, W_K, W_V$. Previous consistency generation [55, 71] is based on concatenated attention via a simple batch image generation, where the first sample in the batch serves as a reference for the $i$-th sample generation. We denote the latent feature for $i$-th sample as $X_i$. Instead of solely depending on its own content, concatenated attention suggests

$$X'_{\texttt{CAT}} = \text{Attention}\left(Q_i, [K_1; K_i], [V_1; V_i]\right), \tag{2}$$

where $Q_i = X_i W_Q$, $K_i = X_i W_K$, and $V_i = X_i W_V$.

### 3.2 Reference feature guidance

To illustrate why concatenated attention can help boost consistency between generated samples with reference samples, we can reformulate eq. (2) as the following (Proof in appendix C)

$$X'_{\texttt{CAT}} = C \odot \text{Attention}\left(Q_i, K_1, V_1\right) + (\mathbf{1} - C) \odot \text{Attention}\left(Q_i, K_i, V_i\right) \tag{3}$$

where $C$ is a rank-1 matrix of the same size as the attention output, $\odot$ is the point-wise multiplication and $\mathbf{1}$ is an all-ones matrix.

Equation (3) depicts that the concatenated attention is a linear interpolation between the output $X'$ without concatenated attention in eq. (1) and cross-attention between the $i$-th image $X_i$ and the reference image $X_1$, while the coefficient matrix $C$ is determined by the synthesized content $X_i$ and the reference content $X_1$. Before we further improve concatenated attention, we first discuss two related questions for eq. (3). **Is linear interpolation a necessity?** It may be tempting to highlight the role of the second cross-attention term naively while keeping the weights for the first term unchanged, such as $\text{Attention}\left(Q_i, K_1, V_1\right) + \text{Attention}\left(Q_i, K_i, V_i\right)$. However, we find that naively breaking the linear interpolation disrupts image generation. In fact, we can interpret concatenated attention in eq. (3) as applying extra guidance on the original self-attention output

$$X'_{\texttt{CAT}} = \text{Attention}\left(Q_i, K_i, V_i\right) + C \odot \left(\text{Attention}\left(Q_i, K_1, V_1\right) - \text{Attention}\left(Q_i, K_i, V_i\right)\right) \tag{4}$$

https://doi.org/10.52202/079017-1058

which resembles the form of guidance used in diffusion literature [53, 22], such as classifier-free guidance [21]. Notably, the linear interpolation helps keep the attention output $X'_{\texttt{CAT}}$ norm close to self attention output $X'$; otherwise, arbitrary weights would pose a training and inference discrepancy and degrade the generation quality. However, different from various guidance methods used in the diffusion literature, the guidance weights in eq. (4) are constants determined by latent features $X_i$ and reference context $X_1$ and have no user control. Therefore we question **Is constant $C$ matrix coefficient is a necessity?** As an attempt to bypass the rigid form of concatenated attention, we propose a simple and flexible approach named *Reference feature guidance* (`RFG`) (see Fig. 2)

$$X'_{\texttt{RFG}} = c \cdot \text{Attention}\left(Q_i, K_1, V_1\right) + (1 - c) \cdot \text{Attention}\left(Q_i, K_i, V_i\right), \tag{5}$$

where $c$ is a scalar coefficient that controls the strength of the reference image influence.

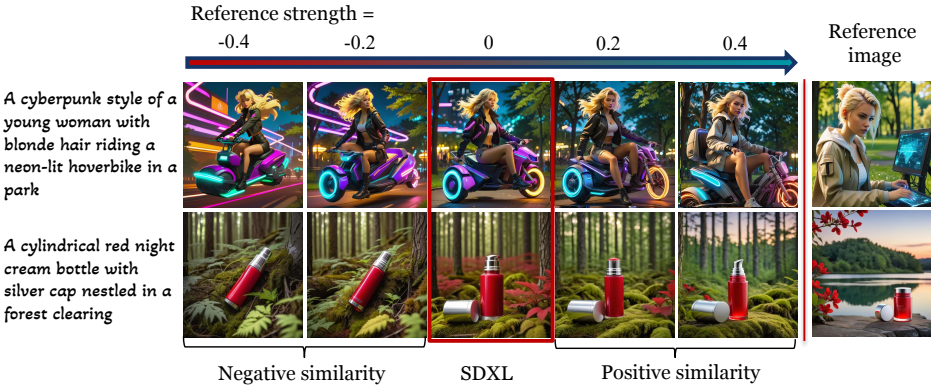


Figure 3: We allow flexible control over the reference effect through a reference strength coefficient.

While most previous methods for consistent generation, including feature combination [30, 71] and injection [55], employ concatenated attention (2), our `RFG` offers several advantages. First, it grants users greater control over the extent of influence from the reference image, as illustrated in Fig. 3. Second, this flexibility proves especially beneficial in a novel application: blending features from multiple reference images. Our method allows users to selectively determine the influence of each reference image. We have observed that the most harmonious blending often results from varying the strength of each reference, rather than maintaining equal strength across all images. Third, by enabling negative coefficients, we find that our method can simulate a concept suppression effect, meaning it generates images that are dissimilar to a reference image. Moreover, it allows for the injection of reference image features into video generation slightly to reduce flickering, while the concatenated attention keeps the video completely static. Finally, our approach is versatile on network architecture as it applies not only to UNet-based models but also to transformer-based diffusion models, such as FLUX model (see appendix D.1).

Therefore, we introduce `RefDrop`, a training-free approach to flexibly control consistency generation, which replaces the self-attention blocks in the diffusion model with `RFG`. For Video Diffusion Models (VDM) [5, 15, 19], we modify *every spatial* self-attention layers to bolster temporal consistency.

## 4 Experiments

We conduct experiments to show that `RefDrop` can help control consistency in two important tasks: image generation and video generation.

### 4.1 Controllable consistency in image generation

We use a fine-tuned SDXL of higher quality, ProtoVision-XL, as the base model for our experiments. For simplicity, we will refer to it as

Table 1: Comparison of Controllable Consistent Image Generation Methods. 'Training-free' indicates no encoder training or diffusion model fine-tuning is needed. 'Single ref.' means the method can operate with only one reference image.

| Name | Training free | Concept suppression | Single ref. |
|---|---|---|---|
| IP-Adapter [67] | ✗ | ✓ | ✓ |
| Consistory [55] | ✓ | ✗ | ✓ |
| Chosen one [1] | ✗ | ✗ | ✗ |
| ELITE [61] | ✗ | ✗ | ✓ |
| BLIPD [33] | ✗ | ✗ | ✓ |
| Ours | ✓ | ✓ | ✓ |


https://doi.org/10.52202/079017-1058

SDXL hereafter. We have replaced all the self-attention layers in SDXL with RFG, using the first sample in the batch as the reference image.

**Evaluation baselines** In this section, we compare RefDrop with several baseline approaches: (1) SDXL [43] without any modifications to its architecture; (2) Ref-ControlNet [1]; (3) encoder-based methods, such as IP-Adapter [67] and BLIPD [33]. For encoder-based methods, we initially generate a reference image using SDXL and then utilize this image as input. Additionally, we present a comparison of several other methods in Table 1.

### 4.1.1 Consistent image generation


Figure 4: The reference image for all methods is framed in red. Our method tends to produce more consistent hairstyles, and facial features compared to IP-Adapter, Ref-ControlNet and BLIPD, and our generation has diverse spatial layout. The visual quality of BLIPD is not comparable, as it utilizes SD1.5 [48] as its base model.

For this task, we use $c \in [0.3, 0.4]$ for our method. However, applying RFG to all the self-attention blocks can lead to the leakage of spatial layout and background from the reference image, causing the generated objects to have quite similar poses and backgrounds. To address these issues, we introduce two techniques: excluding the first upsampling block and applying the subject mask.

**Excluding the first upsampling block** SDXL UNet consists of 4 downsampling blocks, 1 middle block, and 6 upsampling blocks. Through an ablation study, we found that the first upsampling block predominantly influences the spatial layout. As shown in Fig. 5, excluding this block from the modified attention blocks allows for recovering diverse object poses. To the best of our knowledge, this is the first work to use this method for mitigating spatial layout leakage in consistent image generation. Consistory [55] also proposes two techniques to enhance layout diversity: using vanilla query features and self-attention dropout. In comparison, our approach is simpler and more straightforward.

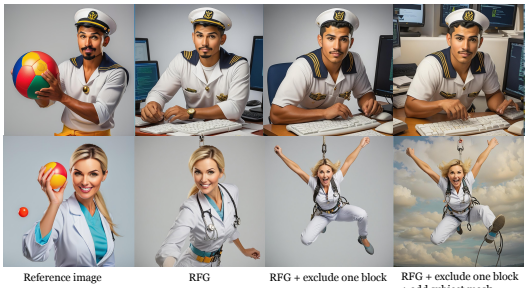


Figure 5: Excluding one block from applying RFG solves the spatial layout leakage issue. Adding subject mask solves the background leakage issue.

**Applying the subject mask** We use Grounded SAM [46] to extract the object mask from the generated reference image by prompting the object name, such as "Guinea pig" or "human." The

[1] https://github.com/Mikubill/sd-webui-controlnet/discussions/1236

mask is then downsampled to match the latent feature resolution of the SDXL UNet. The resulting masked RFG is defined as follows

$$X'_{\texttt{RFG}} = cM \odot \text{Attention}\left(Q_i, K_1, V_1\right) + (\mathbf{1} - cM) \odot \text{Attention}\left(Q_i, K_i, V_i\right), \tag{6}$$

where the mask $M$ ensures that guidance is restricted to the masked area. Note that, our masked RFG (6) does not modify the attention operation itself but only adjusts the coefficients, making it memory efficient and straightforward to implement.

We show qualitative results in Fig. 4. IP-Adapter establishes a strong baseline, especially on single subject consistent generation. However, it suffers from similar spatial layout, and requires additional computational resources and data for training the image encoder compared to our approach. While Reference-only ControlNet performs well on simple subjects, such as cartoon characters, it struggles to generate humans. It is likely to produces humans with distorted eyes and bodies. BLIPD underperforms in terms of both visual quality and consistency relative to RefDrop. For **multi-subject** consistent generation, we find RefDrop can straightforwardly work for semantically different objects even without using separate subject masks. This observation aligns with ConsiStory [55]. Further comparative results are available in Figs. 13 and 16.

#### 4.1.2 Blend features from multiple images

RefDrop also supports the use of multiple reference images. In our implementation, we designate the first $N$ images in a batch as reference images. Features from these reference images are then incorporated into the subsequent images within the same batch through every self-attention layer. Formally, the extended RFG with multiple references is defined as

$$X'_{\texttt{RFG}} = \sum_{j=1}^{N} c_j \cdot \text{Attention}(Q_i, K_j, V_j) + \left(1 - \sum_{j=1}^{N} c_j\right) \cdot \text{Attention}(Q_i, K_i, V_i), \tag{7}$$

for certain $i > N$. Here, the attention mechanism ensures that the $i$-th image in the batch receives features from the first $1 \sim N$ reference images. In practice, we use $c_j \in [0.2, 0.4]$ for any $j = 1, \ldots, N$ in our method. We demonstrate the capability of RefDrop to seamlessly blend distinct semantic features from two reference objects into a new object in Figs. 6, 18 and 19. This task proves challenging when relying solely on prompt engineering. We attempted to achieve this task with SDXL using text prompts. For instance, if we aim to merge two objects, $\alpha$ and $\beta$, we might use prompts like "an $\alpha$-like $\beta$" or "a $\beta$ in the style of $\alpha$." However, with such text prompts, SDXL either ignores the similarity with one of the reference images or frequently produces multiple objects instead of a single and cohesive entity. In Fig. 20, we show that RefDrop can blend *three* distinct subjects: a dwarf, Black Widow, and Winnie the Pooh, encompassing a range of mythological being, human, and animal.


Figure 6: **Multiple Reference Images:** The reference images are highlighted with a red frame, and the third image in each set is the resultant blended image. RefDrop effectively assimilates features from the distinct reference images into a single and cohesive entity, demonstrating robust feature integration capability.

#### 4.1.3 Diverse image generation

Our method offers substantial flexibility in parameter tuning, enabling diverse image generation by setting the coefficient $c$ to a negative value. This feature is particularly valuable in addressing overfitting issues in image generation. For instance, when using SDXL to generate Middle Eastern faces, the output frequently includes similar headscarves, faces and outfits, as illustrated on the left side of Fig. 8.

Figure 7: Negative reference images for examples in Fig. 8.

In this task, we use $c = -0.3$ for our method. We present a qualitative comparison in Fig. 8, with negative reference images displayed in Fig. 7. Upon comparing our method with IP-Adapter, we


https://doi.org/10.52202/079017-1058

note that IP-Adapter may not adhere as closely to the text prompt. We attribute this to IP-Adapter's modification of cross-attention, which can impact text alignment. In contrast, our method focuses on modifying self-attention, thereby preserving the integrity of cross-attention and ensuring more accurate text alignment. We show additional quantitative results in Fig. 15.


Figure 8: Diverse image generation: Our method enhances diversity in outfits, hairstyles, and facial features, all while ensuring accurate text alignment. For example, while SDXL frequently generates headscarves in the first scenario and beige-colored clothes in the second, `RefDrop` can vary the presence of headscarves in the left example and produce clothing in different colors in the right example. Conversely, although IP-Adapter can create even more diverse images, it often fails to adhere to the style and human activity instructions in the text prompts. Additionally, it often produces overly small persons that lack detail.

### 4.2 Improving temporal-consistency in video generation

Not only can we apply `RFG` to T2I generation, but it also effectively stabilizes video generation, where flickering commonly degrades quality. This section shows that using the first generated frame as a reference can greatly improve video generation. By injecting its features into the spatial self-attention layers of subsequent frames with a reference strength of $c = 0.2$, we significantly stabilize these frames and enhance the temporal consistency of VDM.

We employ `SVD-img2vid-xt-1-1` as our I2V base model. Technically, our approach is compatible with any VDM, but we choose SVD as it is the best open-source model available. Although this model usually produces consistent videos from visually perfect images, we have noted that minor, often imperceptible flaws in the input images can significantly degrade the quality of the generated videos. Our method effectively stabilizes video quality in these scenarios.

#### 4.2.1 Temporal-consistent video generation


Figure 9: Comparison of training free techniques to improve temporal consistency in video generation.

In this part, we use the SDXL model to generate an image from a prompt and then pass this image to the SVD model. For **evaluation baselines**, we compare `RefDrop` with several training-free methods: unmodified SVD, Cross-Frame Attention [29], Concatenated Attention [62], and Temporal Low Pass Frequency Filter (LPFF) [69]. Temporal LPFF is arguably superior to Spatial-Temporal LPFF by Wu et al. [63], which shows Spatial-Temporal LPFF can result in blurry frames. We evaluate the Temporal LPFF using a fast sampling method that avoids the computationally intensive process of iteratively performing backward and forward diffusion at each denoising step.

The visualization results are displayed in Fig. 9. We observe that both Cross-Frame Attention and Concatenated Attention result in completely static videos, whereas LPFF shows minimal improvement. Our method proves to be the most effective in preventing flickering while preserving motion.

### 4.2.2 Stabilizing personalized video generation

Finally, we explore the application of `RefDrop` to personalized video generation. Inspired by Ku et al. [31], starting with an image of a person, we use InstantID [59] to generate a personalized initial frame. This frame is then fed into SVD to create a short video. However, we observe that using the output from InstantID for SVD generation leads to a significantly higher failure rate compared to using the initial frame generated by SDXL. We attribute this increased failure rate to InstantID's propensity for producing images with more flaws, such as overly saturated colors, and distorted limbs, highlighting the potential demand for `RefDrop` in this task.


Figure 10: By injecting the features of the first frame into the generation of subsequent frames, `RefDrop` reduces flickering and facial distortions. The additional videos can be viewed here.

Several other methods are available for personalized video generation, as described in works [60, 28, 38, 24]. Our method, which is designed to enhance temporal consistency, can be integrated with some of these existing approaches. For example, in the case of Magic-Me [38], our attention mechanism `RFG` can be incorporated into their AnimateDiff [18] backbone. For the evaluation in this section, we primarily focus on comparisons with naive SVD generation, as it directly relates to our goal of enhancing temporal consistency.

We present such comparison between our `RefDrop` enhanced generation to the naive SVD generation in Fig. 10. `RefDrop` effectively preserves identity during video generation, offering improvements similar to those achieved by increasing the CFG. However, unlike increasing CFG, which often results in over-saturation of videos, our approach does not produce such artifacts. We present additional automatic metrics in Table 3 to show that `RefDrop` can enhance the quality of the generated videos.

## 5 Human evaluation

We conducted a human evaluation study using Google Forms. Our survey is structured into three distinct categories: 1) Consistent Image Generation, 2) Diverse Image Generation, and 3) Personalized Video Generation. Initially, we utilized ChatGPT to generate text prompts, then processed approximately 100 small tasks per category using both baseline methods and our approach. From these, we randomly selected 10 sets for evaluation. In the first two categories, participants assessed the


https://doi.org/10.52202/079017-1058

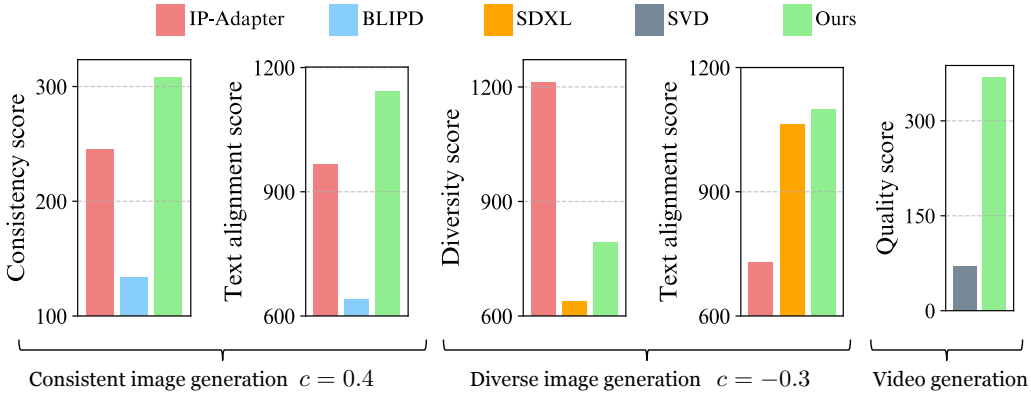


Figure 11: A higher score indicates a superior result. In the category of consistent image generation, participants showed a preference for `RefDrop`, with IP-Adapter ranking slightly behind. In diverse image generation, while IP-Adapter was favored for its variety, it significantly compromised text alignment. Conversely, `RefDrop` maintained a good balance, achieving diversity while preserving text alignment. In personalized video generation, users clearly preferred our approach, demonstrating substantial improvements over the SVD results.

consistency and diversity of the images, as well as text alignment. For the third category, participants were asked to select the video with better quality. The vertical axes in Fig. 11 mean the aggregated scores from all participants. We collected responses from 44 distinct users in total. More details appear in appendix G.

## 6 Conclusion

In this study, we propose a method that effectively uses one or multiple generated images to guide the generation of other images or video frames. Through extensive experiments, our method has proven useful for flexible consistency control in image generation and has improved temporal consistency in video generation. In particular, we show applications in consistent and diverse image generation, feature blending from multiple images, and enhancement of video temporal consistency. Moreover, our approach is versatile on network architecture as it applies not only to UNet-based models but also to transformer-based diffusion models like DiT [42].

Looking ahead, several promising avenues for further research emerge from this study. Firstly, our experiments have not yet explored the use of attention masks; investigating their potential for precise control in image generation presents a compelling opportunity for future work. Another exciting prospect involves enhancing our method to accept clean reference images as input, similar to the IP-Adapter and other image personalization techniques. Achieving this capability would represent a significant advancement, particularly if coupled with an optimal image inversion method.

## Acknowledgments and Disclosure of Funding

We thank Yuval Atzmon and the anonymous reviewers for their valuable feedback on this paper. Jiaojiao Fan, Haotian Xue, and Yongxin Chen are supported in part by grants NSF CAREER ECCS-1942523, and NSF FRR-2409016.

## References

[1] Omri Avrahami, Amir Hertz, Yael Vinker, Moab Arar, Shlomi Fruchter, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. The chosen one: Consistent characters in text-to-image diffusion models. *arXiv preprint arXiv:2311.10093*, 2023.

[2] Omri Avrahami, Rinon Gal, Gal Chechik, Ohad Fried, Dani Lischinski, Arash Vahdat, and Weili Nie. Diffuhaul: A training-free method for object dragging in images. *arXiv preprint arXiv:2406.01594*, 2024.

[3] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, et al. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.

[4] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Yuanzhen Li, Tomer Michaeli, et al. Lumiere: A space-time diffusion model for video generation. *arXiv preprint arXiv:2401.12945*, 2024.

[5] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets, 2023.

[6] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22563–22575, 2023.

[7] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. URL https://openai.com/research/video-generation-models-as-world-simulators.

[8] Di Chang, Yichun Shi, Quankai Gao, Jessica Fu, Hongyi Xu, Guoxian Song, Qing Yan, Xiao Yang, and Mohammad Soleymani. Magicdance: Realistic human dance video generation with motions & facial expressions transfer. *arXiv preprint arXiv:2311.12052*, 2023.

[9] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter1: Open diffusion models for high-quality video generation, 2023.

[10] Xuweiyi Chen, Tian Xia, and Sihan Xu. UniCtrl: Improving the spatiotemporal consistency of Text-to-Video diffusion models via Training-Free unified attention control. March 2024.

[11] Yuren Cong, Mengmeng Xu, Christian Simon, Shoufa Chen, Jiawei Ren, Yanping Xie, Juan-Manuel Perez-Rua, Bodo Rosenhahn, Tao Xiang, and Sen He. Flatten: optical flow-guided attention for consistent text-to-video editing. *arXiv preprint arXiv:2310.05922*, 2023.

[12] Zhangyin Feng, Yuchen Ren, Xinmiao Yu, Xiaocheng Feng, Duyu Tang, Shuming Shi, and Bing Qin. Improved visual story generation with adaptive context modeling. *arXiv preprint arXiv:2305.16811*, 2023.

[13] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data, 2023.

[14] 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.

[15] Songwei Ge, Seungjun Nah, Guilin Liu, Tyler Poon, Andrew Tao, Bryan Catanzaro, David Jacobs, Jia-Bin Huang, Ming-Yu Liu, and Yogesh Balaji. Preserve your own correlation: A noise prior for video diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22930–22941, 2023.

[16] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373*, 2023.

[17] Rohit Girdhar, Mannat Singh, Andrew Brown, Quentin Duval, Samaneh Azadi, Sai Saketh Rambhatla, Akbar Shah, Xi Yin, Devi Parikh, and Ishan Misra. Emu video: Factorizing text-to-video generation by explicit image conditioning, 2023.

[18] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

 https://doi.org/10.52202/079017-1058

[19] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation. *arXiv preprint arXiv:2211.13221*, 2022.

[20] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.

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

[22] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022.

[23] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022.

[24] Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, and Liefeng Bo. Animate anyone: Consistent and controllable Image-to-Video synthesis for character animation. November 2023.

[25] Zehuan Huang, Hongxing Fan, Lipeng Wang, and Lu Sheng. From parts to whole: A unified reference framework for controllable human image generation. *arXiv preprint arXiv:2404.15267*, 2024.

[26] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. URL https://doi.org/10.5281/zenodo.5143773. If you use this software, please cite it as below.

[27] Hyeonho Jeong, Gihyun Kwon, and Jong Chul Ye. Zero-shot generation of coherent storybook from plain text story using diffusion models. *arXiv preprint arXiv:2302.03900*, 2023.

[28] Yuming Jiang, Tianxing Wu, Shuai Yang, Chenyang Si, Dahua Lin, Yu Qiao, Chen Change Loy, and Ziwei Liu. Videobooth: Diffusion-based video generation with image prompts. *arXiv preprint arXiv:2312.00777*, 2023.

[29] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15954–15964, 2023.

[30] Chanran Kim, Jeongin Lee, Shichang Joung, Bongmo Kim, and Yeul-Min Baek. Instantfamily: Masked attention for zero-shot multi-id image generation, 2024.

[31] Max Ku, Cong Wei, Weiming Ren, Huan Yang, and Wenhu Chen. Anyv2v: A plug-and-play framework for any video-to-video editing tasks. *arXiv preprint arXiv:2403.14468*, 2024.

[32] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1931–1941, 2023.

[33] Dongxu Li, Junnan Li, and Steven Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems*, 36, 2024.

[34] 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.

[35] Chang Liu, Haoning Wu, Yujie Zhong, Xiaoyun Zhang, and Weidi Xie. Intelligent grimm–open-ended visual storytelling via latent diffusion models. *arXiv preprint arXiv:2306.00973*, 2023.

[36] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. *arXiv preprint arXiv:2310.11440*, 2023.

[37] Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. *Advances in Neural Information Processing Systems*, 36, 2024.

[38] Ze Ma, Daquan Zhou, Chun-Hsiao Yeh, Xue-She Wang, Xiuyu Li, Huanrui Yang, Zhen Dong, Kurt Keutzer, and Jiashi Feng. Magic-me: Identity-specific video customized diffusion. *arXiv preprint arXiv:2402.09368*, 2024.

[39] Adyasha Maharana, Darryl Hannan, and Mohit Bansal. Storydall-e: Adapting pretrained text-to-image transformers for story continuation. In *European Conference on Computer Vision*, pp. 70–87. Springer, 2022.

[40] Xichen Pan, Li Dong, Shaohan Huang, Zhiliang Peng, Wenhu Chen, and Furu Wei. Kosmos-g: Generating images in context with multimodal large language models. *arXiv preprint arXiv:2310.02992*, 2023.

[41] Xichen Pan, Pengda Qin, Yuhong Li, Hui Xue, and Wenhu Chen. Synthesizing coherent story with auto-regressive latent diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2920–2930, 2024.

[42] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

[43] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

[44] Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[45] Tanzila Rahman, Hsin-Ying Lee, Jian Ren, Sergey Tulyakov, Shweta Mahajan, and Leonid Sigal. Make-a-story: Visual memory conditioned consistent story generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2493–2502, 2023.

[46] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024.

[47] Weiming Ren, Harry Yang, Ge Zhang, Cong Wei, Xinrun Du, Stephen Huang, and Wenhu Chen. Consisti2v: Enhancing visual consistency for image-to-video generation. *arXiv preprint arXiv:2402.04324*, 2024.

[48] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022.

[49] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pp. 234–241. Springer International Publishing, 2015.

[50] 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*, pp. 22500–22510, 2023.

[51] 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.

 https://doi.org/10.52202/079017-1058

[52] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

[53] Jiaming Song, Qinsheng Zhang, Hongxu Yin, Morteza Mardani, Ming-Yu Liu, Jan Kautz, Yongxin Chen, and Arash Vahdat. Loss-guided diffusion models for plug-and-play controllable generation. In *International Conference on Machine Learning*, pp. 32483–32498. PMLR, 2023.

[54] Yoad Tewel, Rinon Gal, Gal Chechik, and Yuval Atzmon. Key-locked rank one editing for text-to-image personalization. In *ACM SIGGRAPH 2023 Conference Proceedings*, pp. 1–11, 2023.

[55] Yoad Tewel, Omri Kaduri, Rinon Gal, Yoni Kasten, Lior Wolf, Gal Chechik, and Yuval Atzmon. Training-free consistent text-to-image generation. *arXiv preprint arXiv:2402.03286*, 2024.

[56] Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. Emo: Emote portrait alive-generating expressive portrait videos with audio2video diffusion model under weak conditions. *arXiv preprint arXiv:2402.17485*, 2024.

[57] Ashish Vaswani, Noam M Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Adv. Neural Inf. Process. Syst.*, pp. 5998–6008, June 2017.

[58] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, Steven Liu, William Berman, Yiyi Xu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. URL https://github.com/huggingface/diffusers.

[59] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, and Anthony Chen. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024.

[60] Zhao Wang, Aoxue Li, Enze Xie, Lingting Zhu, Yong Guo, Qi Dou, and Zhenguo Li. Customvideo: Customizing text-to-video generation with multiple subjects. *arXiv preprint arXiv:2401.09962*, 2024.

[61] 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15943–15953, 2023.

[62] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7623–7633, 2023.

[63] Tianxing Wu, Chenyang Si, Yuming Jiang, Ziqi Huang, and Ziwei Liu. Freeinit: Bridging initialization gap in video diffusion models. *arXiv preprint arXiv:2312.07537*, 2023.

[64] 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.

[65] Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Rerender a video: Zero-shot text-guided video-to-video translation. In *SIGGRAPH Asia 2023 Conference Papers*, pp. 1–11, 2023.

[66] Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Fresco: Spatial-temporal correspondence for zero-shot video translation. *arXiv preprint arXiv:2403.12962*, 2024.

[67] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. 2023.

[68] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

[69] Yabo Zhang, Yuxiang Wei, Xianhui Lin, Zheng Hui, Peiran Ren, Xuansong Xie, Xiangyang Ji, and Wangmeng Zuo. Videoelevator: Elevating video generation quality with versatile text-to-image diffusion models. *arXiv preprint arXiv:2403.05438*, 2024.

[70] Youyuan Zhang, Xuan Ju, and James J Clark. Fastvideoedit: Leveraging consistency models for efficient text-to-video editing. *arXiv preprint arXiv:2403.06269*, 2024.

[71] Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation, 2024.

 https://doi.org/10.52202/079017-1058

# A Limitation

- For consistent image generation, our model sometimes struggles to accurately replicate specific objects like lotion bottles, often hallucinating instead. It also occasionally fails to replicate the exact hairstyle and outfit from the reference person or animal.
- For diverse image generation, our method currently cannot precisely control aspects of diversity, such as generating diverse subjects but not styles, or vice versa.
- For blending multiple images, the coefficient needs careful tuning to achieve a good balance among multiple reference images, and it heavily depends on the specific case.
- For improving video temporal consistency, although our method successfully generates videos with temporal consistency, the generated videos contain less motion. A straightforward way to alleviate this issue is to adjust the RFG coefficient depending on the case. To fully address this issue, we believe introducing additional training-based modules, for example, a motion guidance module proposed in StoryDiffusion [71], could provide stronger motion guidance. Alternatively, we could add motion frames as conditional guidance, as done in EMO [56].
- Finally, when applying our method to models beyond SDXL or SVD, we find that the coefficient parameter $c$ may require additional tuning. Additionally, for the technique we proposed to address spatial layout leakage, an ablation study is necessary to identify which layer primarily governs spatial layout generation, as this layer may vary across different models.

# B Broader impacts

The broader impacts of advancements in consistent character generation and personalized video generation extend across multiple domains, notably enhancing both creative and technological landscapes. In the media and entertainment industries, for instance, these methods can revolutionize character design, fostering more reliable representations. For media and advertising, it can enhance the design of key frames in advertisements or movie videos, allowing companies to create more visually compelling content. Individual artists may leverage this technique to blend multiple images, potentially sparking new creative inspirations. RefDrop could be used to improve temporal consistency in short videos, which may be particularly valuable for social media platforms seeking to enhance AI-generated video content. However, there is also a potential risk associated with our method, as it could be used to create fake profiles, highlighting the need for careful consideration of its applications.

# C Relationship to Concatenated attention

## C.1 Mathematical equivalence

We firstly point out that our RFG framework can recover the concatenated attention (2) by replacing the scalar coefficient in (5) to be a rank-1 matrix. Then we delve into details. We begin by noting the dimensions of the attention maps:

$$\text{Softmax}\left(\frac{Q_i K_1^\top}{\sqrt{d}}\right) \in \mathbb{R}^{L\times L}, \tag{8}$$

$$\text{Softmax}\left(\frac{Q_i K_i^\top}{\sqrt{d}}\right) \in \mathbb{R}^{L\times L}, \tag{9}$$

$$\text{Softmax}\left(\frac{Q_i [K_1; K_i]^\top}{\sqrt{d}}\right) \in \mathbb{R}^{L\times 2L}, \tag{10}$$


Eq. (11)-(13) Concat attn (2)

Figure 12: Concatenated attention is our special case.

where $L$ is the sequence length of the input hidden feature $X$. We further define $\mathbf{1}_d$ as an all-ones vector of dimension $d$, and $\mathbf{1}$ as an all-ones matrix, sized appropriately to ensure the validity of the operations it is involved in. To recover concatenated attention (2), we extend the scalar $c$ from (5) to a rank-1 weight matrix:

$$C = \mathbf{c} \otimes \mathbf{1}_{d_v}, \tag{11}$$

where all columns in $C \in \mathbb{R}^{L \times d_v}$ are identical, represented by the vector $\mathbf{c} \in \mathbb{R}^{d_v}$, and $d_v$ is the feature dimension of the value $V$. We then transform the scalar dot product into a matrix element-wise product $\odot$, allowing RFG to be expressed with this matrix coefficient as:

$$X'_{\texttt{RFG}} = C \odot \left( \text{Softmax} \left( \frac{Q_i K_1^\top}{\sqrt{d}} \right) V_1 \right) + (\mathbf{1} - C) \odot \left( \text{Softmax} \left( \frac{Q_i K_i^\top}{\sqrt{d}} \right) V_i \right). \tag{12}$$

Denote $./$ and exp as the element-wise division and exponential operation respectively. By setting

$$\mathbf{c} = \left( \exp \left( \frac{Q_i K_1^\top}{\sqrt{d}} \right) \mathbf{1}_L \right) ./ \left( \exp \left( \frac{Q_i [K_1; K_i]^\top}{\sqrt{d}} \right) \mathbf{1}_{2L} \right), \tag{13}$$

$X'_{\texttt{RFG}}$ can recover the concatenated attention [62]

$$X'_{\texttt{CAT}} = \text{Softmax} \left( \frac{Q_i [K_1; K_i]^\top}{\sqrt{d}} \right) [V_1; V_i] = \text{Softmax} \left( \frac{[Q_i K_1^\top; Q_i K_i^\top]}{\sqrt{d}} \right) [V_1; V_i]. \tag{14}$$

The reason for this to hold is simply the normalizing effect of softmax. The softmax operation would normalize each row in the attention map $\frac{Q_i [K_1; K_i]^\top}{\sqrt{d}}$ independently, thus the weight matrix to recover the concat attention is a rank-1 matrix with different rows.

We want to note that the rank-1 matrix used in the concatenated attention-based method is not explicitly shown because they use Equation (2). One of our contributions is demonstrating that Equation (2) can be equivalently reformulated as Equation (3), where the rank-1 matrix appears as the coefficient. We further simplify Equation (3) by changing the rank-1 matrix to a scalar. We emphasize that in the concatenated attention-based method, the rank-1 matrix is not manually adjustable or defined by the user; it is intrinsically determined by the reference image feature and generated image feature.

We also note that if we only use a scalar as the coefficient, we cannot exactly replicate the concatenated attention method. However, we empirically find that using a coefficient of 0.3∼0.4 can produce a similar effect to the concatenated attention method.

Finally, we can also recover the Cross-Frame attention [29] by setting the coefficient $c = 1$.

### C.2 Visual comparison

In Fig. 13, we present a visual comparison with concatenated attention for consistent image generation task, using the same random seed and without additional techniques such as excluding one attention block (as proposed by us) and self-attention dropout from Consistory [55]. In practice, we find that RFG with a coefficient of 0.3∼0.4 produces results quite similar to the concatenated attention method used in Tewel et al. [55], Zhou et al. [71].

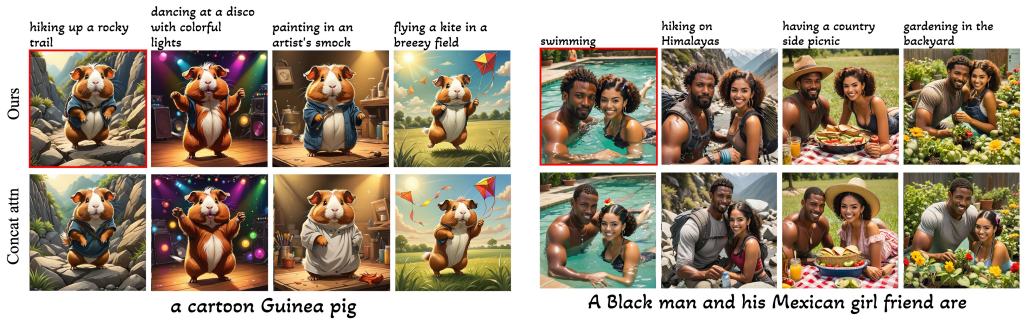


Figure 13: Comparison with Concatenate attention for the consistent image generation task. Both of ours and Concatenated attention methods do not apply subject mask and are applied to all attention blocks here for a fair comparison.


https://doi.org/10.52202/079017-1058

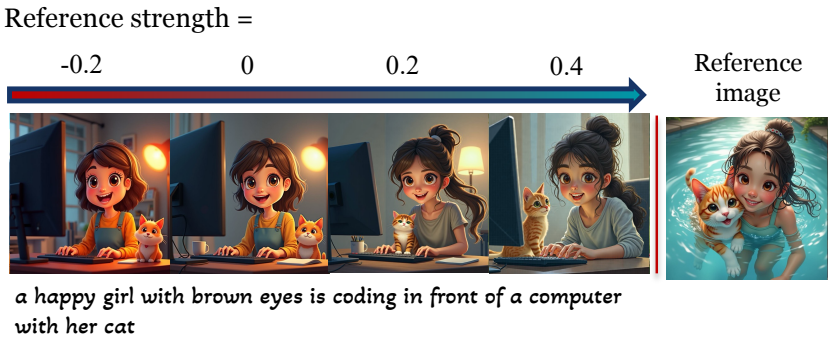


Figure 14: Applying Reference Feature Guidance on the FLUX-dev model. Our method effectively applies reference guidance to the generation.

# D Additional results

## D.1 Applying RFG on transformer-based architecture

We apply Reference Feature Guidance on the most recent capable open-source T2I model, FLUX-dev, developed by Black Forest Lab. We show the effect of Reference Feature Guidance on the FLUX-dev model in Fig. 14. Since FLUX is a transformer-based diffusion model, it does not have separate cross-attention and self-attention layers. we have to do some modification to adapt to their model. Every attention block in FLUX-dev is:

$$X' = \text{Attention}([Q_{\text{img}}; Q_{\text{txt}}], [K_{\text{img}}; K_{\text{txt}}], [V_{\text{img}}; V_{\text{txt}}])$$

where $[\cdot;\cdot]$ denotes the concatenation operation. Similar to (4), we modify the attention blocks in FLUX to accept additional guidance from the reference image feature, assumed to be the first image in the batch:

$$X'_{\texttt{RFG}} = X' + c \cdot (\text{Attention}(Q_{\text{img}\,i}, K_{\text{img}\,1}, V_{\text{img}\,1}) - \text{Attention}(Q_{\text{img}\,i}, K_{\text{img}\,i}, V_{\text{img}\,i})) \tag{15}$$

We add two additional terms multiplied by a guidance scale $c$ on top of the original attention output $X'$. And those additional terms only depend on the image features, ensuring that we do not interfere with the text features. Fig. 14 demonstrates that our method effectively applies reference guidance to the generation.

## D.2 Quantitative results

**Importance of the first upsampling block for SDXL UNet** In Sec. 4.1.1, we introduced a technique to mitigate the spatial layout leakage from the reference image. In Table 2, we show quantitative result to verify that removing the first upsampling block is more effective than removing other blocks here to remove spatial layout leakage. We conducted 11 groups of experiments, where in each group, we excluded one of the 11 blocks of the UNet from applying RFG. And in each group, we generated 20 sets of consistent objects, with each set containing 5 images. This resulted in a total of $11 \times 20 \times 5 = 1100$ images for metric calculation. We then used DreamSim and LPIPS to measure the distance between the generated images and the reference image, and report the mean and standard deviation. Higher values of these metrics should indicate more diverse poses, as these scores tend to favor similar layouts [55]. From Table 2, we can see that excluding the first upsampling block greatly boosts the spatial layout diversity.

**Text-to-Image generation** We present quantitative metrics in Figure 15. Using the OpenCLIP model, `CLIP-ViT-g-14-laion2B`, we measure text-image similarity by averaging CLIP scores [20] across 100 pairs of text prompts and generated images. This measurement is repeated five times using different pairs for each method, and the variability is depicted through error bars. For assessing subject consistency, we utilize DreamSim [13], after processing images to remove backgrounds[2]

[2]We use the Tracer-B7 model in
https://github.com/OPHoperHPO/image-background-remove-tool/?tab=readme-ov-file

Table 2: Comparison of DreamSim and LPIPS distances for excluding different blocks. The Up1 block of SDXL UNet shows the highest values for both metrics, indicating its strong impact on spatial layout diversity.

| Excluded Block | DreamSim to reference image | LPIPS to reference image |
|---|---|---|
| Down1 | $0.2283 \pm 0.0102$ | $0.5484 \pm 0.0052$ |
| Down2 | $0.2261 \pm 0.0118$ | $0.5484 \pm 0.0043$ |
| Down3 | $0.2298 \pm 0.0082$ | $0.5475 \pm 0.0058$ |
| Down4 | $0.2332 \pm 0.0095$ | $0.5551 \pm 0.0072$ |
| Mid | $0.2484 \pm 0.0048$ | $0.5673 \pm 0.0070$ |
| Up1 | $\mathbf{0.3077 \pm 0.0092}$ | $\mathbf{0.5880 \pm 0.0047}$ |
| Up2 | $0.2567 \pm 0.0087$ | $0.5603 \pm 0.0077$ |
| Up3 | $0.2440 \pm 0.0088$ | $0.5580 \pm 0.0032$ |
| Up4 | $0.2441 \pm 0.0079$ | $0.5611 \pm 0.0074$ |
| Up5 | $0.2368 \pm 0.0106$ | $0.5540 \pm 0.0055$ |
| Up6 | $0.2455 \pm 0.0103$ | $0.5639 \pm 0.0075$ |

in order to focus analysis on foreground content. In tasks of diverse image generation, we employ LPIPS to gauge image diversity. We calculate the pairwise DreamSim or LPIPS distance between 400 image pairs per method, repeating these measurements with distinct pairs to ensure robust results, and report these findings with error bars. The measures of consistency and diversity are expressed as one minus the calculated DreamSim or LPIPS distances. These results demonstrate that our method is effectively situated on the Pareto-front, aligning with the human evaluations reported in Figure 11.

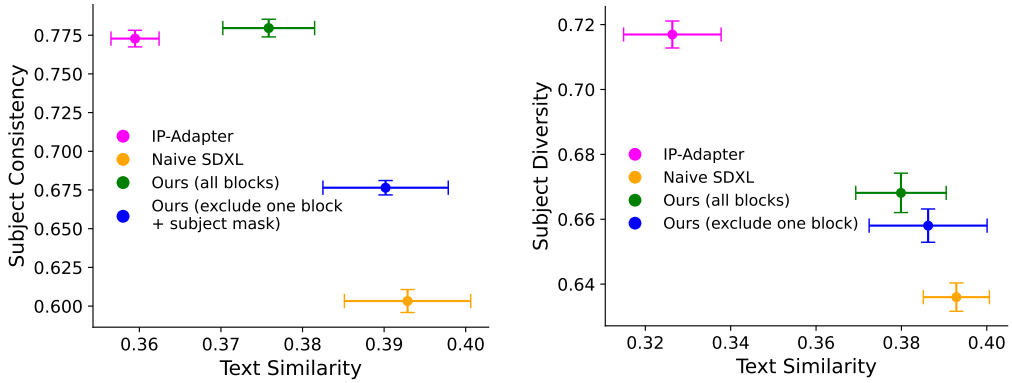


Figure 15: Left: **Consistent Image Generation.** Our method achieves a good balance between text alignment and subject consistency. The techniques of excluding the first upsampling block from being influenced by the reference image (see Sec. 4.1.1) and adding subject masks (see Sec. 4.1.1) help achieve diverse spatial poses and resolve background issues, leading to significant improvements in text alignment. Right: **Diverse Image Generation.** Our approach maintains higher subject diversity with only a slight compromise in text alignment. In contrast, IP-Adapter exhibits the highest subject diversity but suffers from a significant reduction in text alignment. The techniques of excluding the first upsampling block (see Sec. 4.1.1) can also help improve the text alignment here. Error bars represent standard deviation.

**Image-to-Video generation** We present a comparison of automatic metrics for video generation in Table 3. All metrics are designed by EvalCrafter [36]. Following EvalCrafter, we measure the quality of generated videos from four perspectives: overall quality, text alignment, temporal consistency, and motion quality. Specifically, $\text{VQA}_A$ measures the aesthetic score, and $\text{VQA}_T$ evaluates common


https://doi.org/10.52202/079017-1058

distortions such as noise and artifacts. CLIP Score quantifies the similarity between input text prompts and generated videos. For temporal consistency, we use CLIP-Temp to measure semantic consistency between frames, and also calculate face consistency, and warping errors. Finally, the flow score calculates the average optical flow across all video frames. We generated 220 personalized videos using 220 distinct prompts for both SVD and `RefDrop`, utilizing images of four individuals shown in Fig. 10. The prompts included both close-up and distant descriptions. The metrics shown in Table 3 are averaged over these 220 videos. The statistics demonstrate that `RefDrop` reduces unnecessary flickering and improves overall quality. Surprisingly, we find that `RefDrop` not only improves the visual quality but also the text alignment.

Table 3: Comparison of automatic metrics between SVD and `RefDrop` on video generation. An $\uparrow$ symbol indicates that higher values are better, while a $\downarrow$ symbol indicates that lower values are preferable. Our model shows improvements over the SVD base model in overall quality, text alignment, and temporal consistency. The flow score is the only metric where the SVD model scores higher, indicating more motion. However, the SVD model also exhibits greater jittering and flickering, as reflected in its larger warping error. Notably, a static video would register a flow score of zero. This suggests that our generated videos maintain a reasonable level of motion.

| | **Overall quality** | | **Text alignment** | **Temporal Consistency** | | | **Motion** |
|---|---|---|---|---|---|---|---|
| | $\text{VQA}_A \uparrow$ | $\text{VQA}_T \uparrow$ | CLIP score $\uparrow$ | CLIP Temp $\uparrow$ | Face consis. $\uparrow$ | Warping error $\downarrow$ | Flow score $\uparrow$ |
| Ours | **94.27** | **89.91** | **20.84** | **99.91** | **99.46** | **0.0058** | 2.62 |
| SVD | 93.25 | 86.20 | 20.76 | 99.83 | 99.20 | 0.0077 | **5.80** |

### D.3 Qualitative results

**Consistent and Diverse Image Generation:** We give more visualizations for consistent and diverse image generation in Fig. 16 and Fig. 17. We attached images their original quality in `consistent_generation_remove_up1_mask.pdf` and `diverse_generation.pdf` in the supplementary material.

**Blend multiple images:** We show additional blended images using multiple reference images in Fig. 18, Fig. 19, and Fig. 20. In particular, Fig. 18, Fig. 19 utilize two reference images, and Fig. 20 blends three reference images.

**Personalized Video Comparisons:** We show additional comparison in Fig. 21. Moreover, we offer more than 20 original videos in $1024 \times 576$ resolution, accessible via this anonymous external link. On the linked page, the left column displays the video generations of SVD, while the right column features the enhanced SVD results by `RefDrop`.

## E Effect of the coefficient $c$

- **Consistent Image Generation:** More challenging tasks typically require larger coefficients to ensure consistency. For example, generating human figures, which are more complex, requires coefficients between $[0.3, 0.4]$. In contrast, simpler subjects like fluffy toys or cartoon characters may only need a coefficient of $0.2$ to achieve consistent generation.
- **Blend multiple images:** We find that the coefficients for each reference image, typically falling within the range of $[0.2, 0.4]$, perform effectively.
- **Diverse Image Generation:** We recommend using a coefficient of $c = -0.3$. Lower strengths can impair visual quality and may introduce artifacts.
- **Video Consistency:** The coefficient for video consistency requires more nuanced control; A coefficient of $0.2$ generally suffices, and a larger coefficient may make the video totally static. This sufficiency is likely due to the temporal attention component in VDM, which tends to amplify the effects introduced through self-attention.

The effect of reference strength on image generation is in Figs. 3 and 14.

## F Additional implementation details

All experiments were conducted on a single NVIDIA A100 GPU with 80GB of memory. The generation process for a single image using SDXL requires approximately 5 seconds, whereas generating a video using SVD takes about 30 seconds. Additional details on hyper-parameters for both baseline methods and our approach are provided in Table 4.

Table 4: Base model and hyper-parameters.

| | Base model | CFG | Our reference strength | IP-Adapter scale | TLPFF |
|---|---|---|---|---|---|
| Sec. 4.1.1 | Protovision-XL | 5 | 0.3∼0.4 | 0.6 | N/A |
| Sec. 4.1.2 | Protovision-XL | 5 | 0.2∼0.4 | N/A | N/A |
| Sec. 4.1.3 | Protovision-XL | 5 | -0.3 | -0.6 | N/A |
| Sec. 4.2.1 | SVD-img2vid-xt-1-1 | 2.5 | 0.2 | N/A | Gaussian filter Stop frequency = 0.5 |
| Sec. 4.2.2 | SVD-img2vid-xt-1-1 | 2.5 | 0.2 | N/A | N/A |

## G Human evaluation details

The Google Forms survey contains 5 sections, encompassing a total of 50 questions. Instructions and examples are detailed in attached screenshots for each section.

1. **Visual Consistency in Consistent Image Generation**: Participants evaluate visual consistency across four images of the same subject, for example, "Native American sailor" produced by different methods, Methods that maintain character consistency are scored 1; others receive a score of 0. Detailed instructions are provided in Fig. 22.
2. **Text Alignment in Consistent Image Generation**: Respondents assess the alignment of text with the corresponding image for each method, assigning a score from 1 to 3, where a higher score indicates better alignment. Detailed instructions are provided in Fig. 23.
3. **Visual Diversity in Diverse Image Generation**: Like the first section, participants rate the diversity in five images of the same subject across different methods. They assign a score from 1 to 3, with a higher score indicating greater diversity. Detailed instructions are provided in Fig. 24.
4. **Text Alignment in Diverse Image Generation**: This section mirrors Section 2 but in the context of diverse image generation. Participants rate text-image alignment on a scale from 1 to 3. Detailed instructions are provided in Fig. 23.
5. **Personalized Video Quality**: Participants evaluate the quality of videos generated with the same random seed by different methods. Methods that are chosen for higher quality receive a quality score of 1; others receive a score of 0. Detailed instructions are provided in Fig. 25.

We aggregated scores from all sections and display the results in Fig. 11. For the meaning of the scores, i.e. the vertical axes in Fig. 11: we asked the participants to rate the consistency, diversity, etc., according to the rules outlined above. The value on the vertical axis represents the summation of the scores across all participants and questions.

For consistent image generation, we included 5 images per question per method. Four images evaluated subject consistency, and one image evaluated text alignment. There were 10 questions and 3 methods to compare, totaling 150 images. For diverse image generation, we followed a similar approach: 5 images per question per method, 10 questions, and 3 methods, totaling 150 images. For video generation, we included 10 videos and 2 methods, totaling 20 videos. **In total, each participant provided 140 ratings, resulting in 6,160 ratings from 44 participants.**

The results used in the human evaluation did not apply the techniques for mitigating spatial layout and background leakage described in Sec. 4.1.1. Nevertheless, our method is still preferred by the evaluators.


https://doi.org/10.52202/079017-1058

## H Licenses

Pretrained models:

- ProtoVision-XL[3] [43] CreativeML Open RAIL++-M License
- Stable-Video-Diffusion-img2vid-xt-1-1[4] [5] CreativeML Open RAIL++-M License
- FLUX-dev[5] FLUX.1 [dev] Non-Commercial License
- BLIP Diffusion[6] [33] Apache 2.0 License
- IP-Adapter-SDXL[7] [67] Apache 2.0 License
- InstantID[8] [59] Apache 2.0 License

Codebase:

- diffusers 0.25.1 [9] [58] Apache 2.0 License
- EvalCrafter [10] [36] No license found

Metric models:

- OpenCLIP[11] [44, 26] MIT License
- DreamSim[12] [13] MIT License
- LPIPS 1.0[13] [68] BSD-2-Clause license

[3] https://huggingface.co/stablediffusionapi/protovision-xl-high-fidel
[4] https://huggingface.co/stabilityai/stable-video-diffusion-img2vid-xt-1-1
[5] https://huggingface.co/black-forest-labs/FLUX.1-dev
[6] https://huggingface.co/salesforce/blipdiffusion
[7] https://huggingface.co/h94/IP-Adapter/blob/main/sdxl_models/ip-adapter_sdxl.bin
[8] https://huggingface.co/InstantX/InstantID
[9] https://github.com/huggingface/diffusers
[10] https://github.com/EvalCrafter/EvalCrafter
[11] https://huggingface.co/laion/CLIP-ViT-g-14-laion2B-s12B-b42K
[12] https://dreamsim-nights.github.io/
[13] https://lightning.ai/docs/torchmetrics/stable/image/learned_perceptual_image_patch_similarity.html

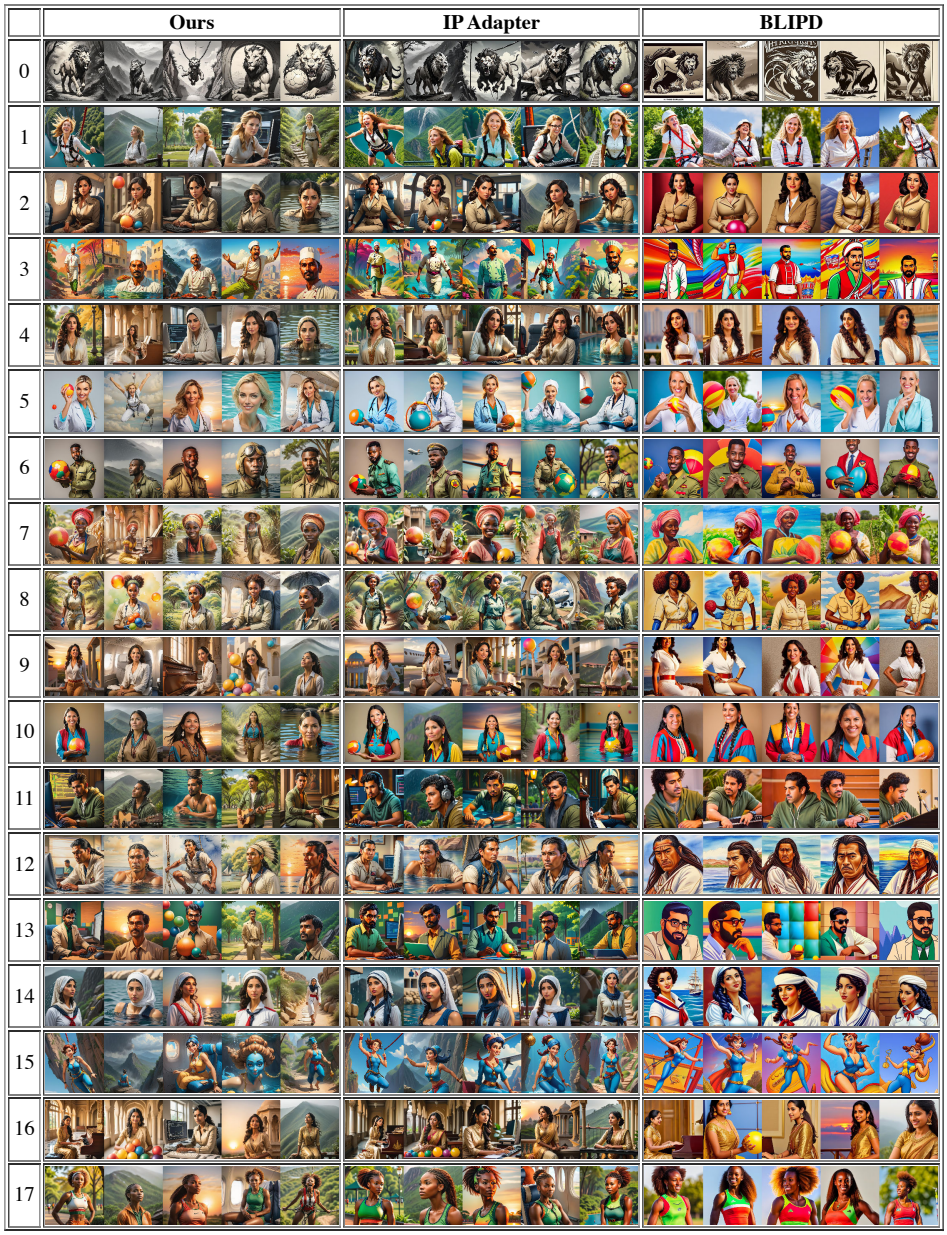


Figure 16: More consistent image generation comparison.


https://doi.org/10.52202/079017-1058

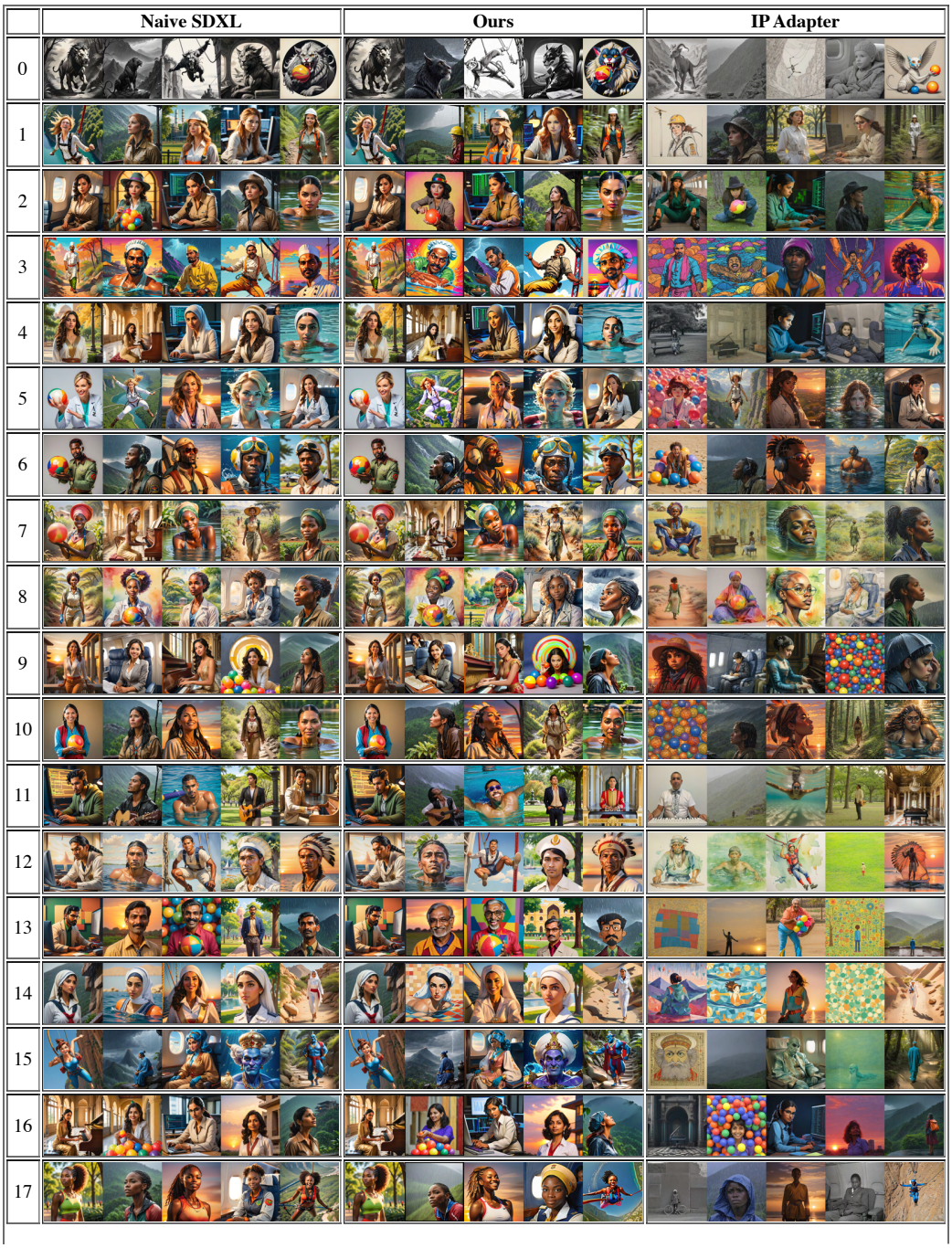


Figure 17: More diverse image generation comparison.

https://doi.org/10.52202/079017-1058


Figure 18: Blending a dog and a cat in various activities: `RefDrop` successfully combines features from two reference images and closely follows the text prompt, whereas SDXL struggles to generate a single cohesive object even with the guidance from the text prompt.

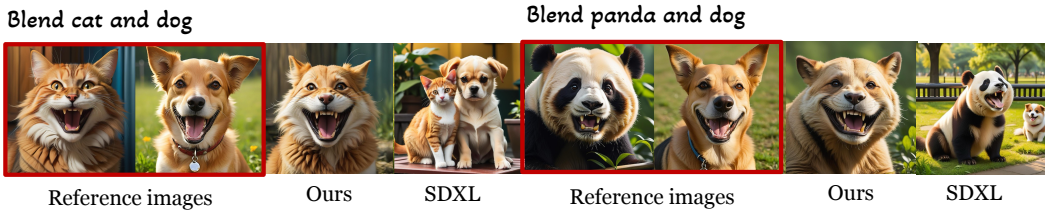


Figure 19: More visualizations for blending two distinct animals. One crucial strategy for our method to effectively blend two objects is to avoid explicitly naming them in the text prompt. We have discovered that using a generic term like "an animal" leads to better results than specifying "a cat-like dog." This trick minimizes the overly strong influence that explicit names can have, facilitating a more effective merger of the two subjects. For SDXL, we use the prompt "a chimera of [animal A] and [animal B]", but it fails to generate a single and cohesive entity.


Figure 20: Blending **three** distinct subjects, we use the same prompt—"a portrait of Winnie the Pooh with red hair and a gray beard"—for both SDXL and `RefDrop`. However, SDXL significantly downplays the features of Winnie the Pooh. In contrast, our approach effectively absorbs the features from the reference images, retaining the dwarf's outfit and beard, Black Widow's red hair, and Winnie's facial structure.


https://doi.org/10.52202/079017-1058


Figure 21: Additional personalized video comparison. The original videos can be viewed here.

https://doi.org/10.52202/079017-1058


Figure 22: The instruction and example for human evaluation.


https://doi.org/10.52202/079017-1058

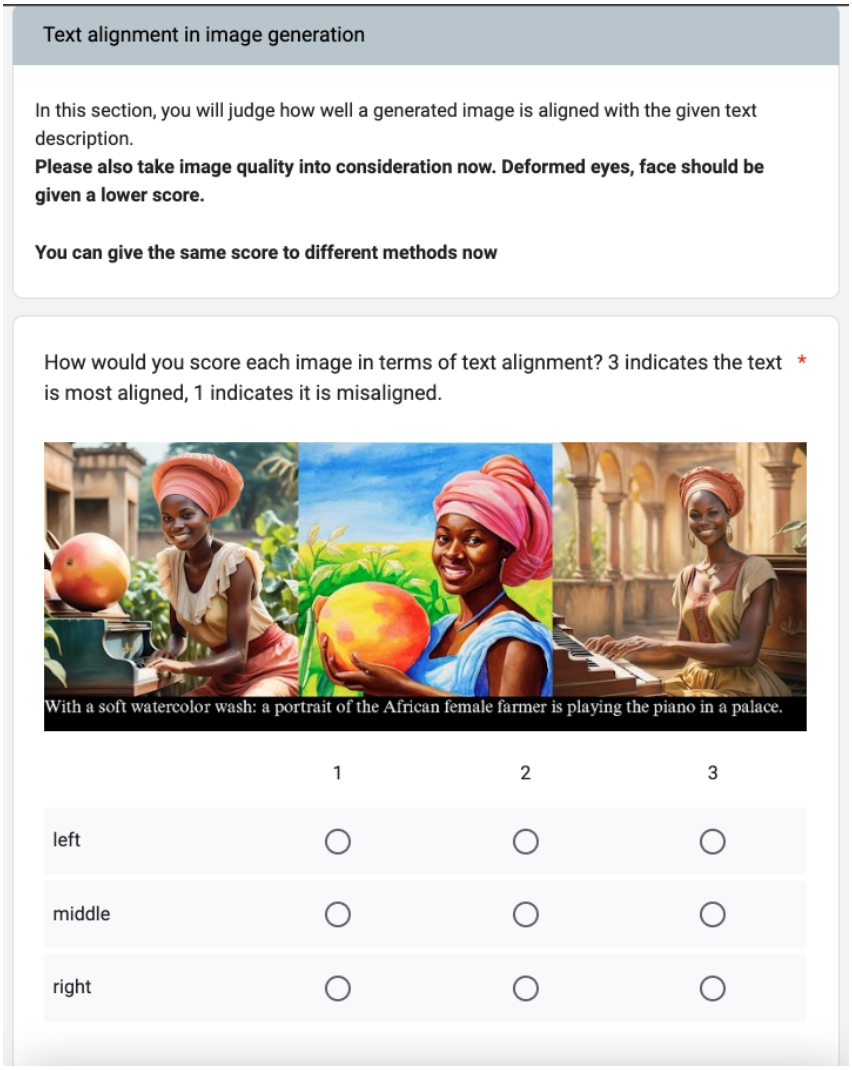


Figure 23: The instruction and example for human evaluation.

Diverse image generation

In this section, you will be asked which group of images are most diverse in terms of

- **human: pose, accessories, hair style, age;**
- **with or without human;**
- **background, camera angle, camera focus length, image style**

For example, if a group of images can generate different accessories, such as eyeglasses, earphones, hats, that it is considered more diverse

In this section, **you need to give a rank, you cannot choose tied ranking**

Which row of images contains the most diverse content? Assign a score of 3 for the most diversity and a score of 1 for the least. *

| | 1 | 2 | 3 |
|---|---|---|---|
| Row 1 | ○ | ○ | ◉ |
| Row 2 | ○ | ◉ | ○ |
| Row 3 | ○ | ○ | ◉ |

Figure 24: The instruction and example for human evaluation.


https://doi.org/10.52202/079017-1058


Figure 25: The instruction and example for human evaluation.

## 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: We made accurate claims in the abstract and introduction.

   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: We have stated our limitation in appendix A.

   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]


https://doi.org/10.52202/079017-1058

Justification: Our theoretical result of the relationship between RFGand concatenated attention is presented in Sec. 3.2, with the proof provided in appendix C.

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: We provide the code of RFG in the supplementary material. It can be directly plugged into SDXL or SVD pipelines in diffusers library.

   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: [No]

Justification: We will open-source the data and code soon.

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: Our appendix F is devoted to experimental details.

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: [Yes]

Justification: We include through statistics in appendix D.2.

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.)


https://doi.org/10.52202/079017-1058

- 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 provide details in appendix F.

   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: When we conduct experiments, we tried our best to cover all races and genders during image or video generation.

   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 discuss Broader Impacts in appendix B.

    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: [NA]

    Justification: We introduce a technique applicable to generative diffusion models. However, we do not release any data or pretrained models at this time.

    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: We provide licenses in appendix H.

    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.


https://doi.org/10.52202/079017-1058

- 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: [NA]

    Justification: We do not release new assets.

    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: [Yes]

    Justification: We discussed the user study details in appendix G.

    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: [No]

    Justification: Our user study questions are simple and short, and we do not foresee potential risks in user study.

    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.