Consistent Story Generation: Unlocking the Potential of Zigzag Sampling

Mingxiao Li
KU Leuven
mingxiao.li@kuleuven.be

Mang Ning
Utrecht University
m.ning@uu.nl

Marie-Francine Moens
KU Leuven
sien.moens@kuleuven.be

Abstract

Text-to-image generation models have made significant progress in producing high-quality images from textual descriptions, yet they continue to struggle with maintaining subject consistency across multiple images, a fundamental requirement for visual storytelling. Existing methods attempt to address this by either fine-tuning models on large-scale story visualization datasets, which is resource-intensive, or by using training-free techniques that share information across generations, which still yield limited success. In this paper, we introduce a novel training-free sampling strategy called Zigzag Sampling with Asymmetric Prompts and Visual Sharing to enhance subject consistency in visual story generation. Our approach proposes a zigzag sampling mechanism that alternates between asymmetric prompting to retain subject characteristics, while a visual sharing module transfers visual cues across generated images to enforce consistency. Experimental results, based on both quantitative metrics and qualitative evaluations, demonstrate that our method significantly outperforms previous approaches in generating coherent and consistent visual stories. The code is available at `https://github.com/Mingxiao-Li/Asymmetry-Zigzag-StoryDiffusion`.

1 Introduction

In recent years, breakthroughs in visual generation techniques have fundamentally transformed the way visual content is created. State-of-the-art image [1–4] and video generation models [5–10] now enable users to produce highly diverse visual outputs with remarkable realism and flexibility. These models can be guided by various forms of control, including bounding boxes [11, 12], object motion trajectories [13, 14], image prompts [5, 4], and even brain signal inputs [15–17], significantly expanding the creative possibilities in both professional and everyday settings.

Despite these impressive advances, challenges remain, particularly in visual storytelling tasks, where multiple pieces of visual content must consistently preserve the identity and key characteristics of subjects across scenes, three examples are presented in Figure 1. A dominant approach to this problem is personalization, which involves learning a subject-specific embedding [18–23]. While effective, this method has notable limitations: the need for per-subject fine-tuning makes it resource-intensive and difficult to scale, and it often leads to overfitting, which can degrade the model's general understanding of text and reduce its generative diversity. To address these limitations, past research has explored tuning-free methods [24–26], which encode subject information using a unified image encoder and leverage large-scale pretraining. These methods eliminate the need for per-subject optimization and offer greater scalability. However, they rely on high-quality, large-scale personalization datasets that are costly to curate and demand substantial computational resources for effective training.

More recently, a line of training-free methods has emerged, targeting the specific challenge of subject-consistent visual story generation without fine-tuning or extensive subject-specific data. The

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

Figure 1: **Visual Story Telling** requires maintaining subject consistency across a sequence of generated images while ensuring that each image faithfully reflects the corresponding prompt. (Images generated using the FLUX model with our proposed method.)

StoryDiffusion model [27] introduces consistent attention by constructing image-level interactions through the random sharing of visual features across different frames. In contrast, the ConsistentStory model [28] aims to preserve subject identity by establishing visual interactions between images more selectively. However, while both methods show promise in maintaining subject consistency, they suffer from notable drawbacks: a reduced ability to follow textual prompts and diminished generative diversity. These issues arise from their direct interference with the diffusion model's generative process. Departing from these strategies, the One-Prompt One-Story model [29] proposes a different approach. First, it fuses all the prompts from a story into a single extended prompt. Then, it generates each image by reweighting the prompt embeddings to reflect the desired focus for each scene. In addition, it improves text-to-image attention by reinforcing subject-relevant information within the prompt itself. Although this model achieves good prompt fidelity, it struggles to maintain subject consistency, particularly when the prompt offers limited detail about the subject.

The above approaches show that directly sharing visual features during the generative process tends to weaken the model's ability to follow textual prompts, while relying solely on prompts to maintain subject consistency proves insufficient. To address both issues, we propose a novel method: **A**symmetry **Z**igzag **S**ampling (AZS). Unlike prior methods that inject subject-specific information into the intermediate representations of each generated image, our approach focuses on incorporating subject information directly into the latent representations of each scene. This distinction allows for more effective control over subject consistency while preserving alignment with textual prompts. As illustrated in Figure 2, our method strategically leverages latent-level visual sharing to improve narrative coherence in multi-scene generation.

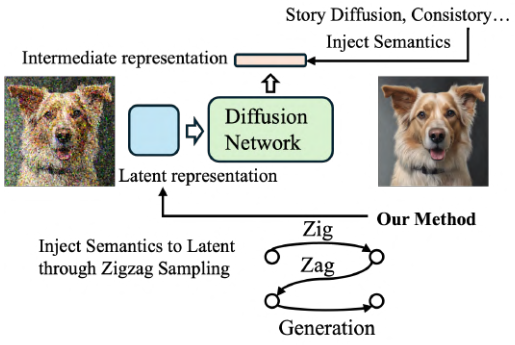

Figure 2: An illustration of Zigzag Sampling, along with a comparison to previous methods in how semantic information is incorporated during image generation.

As shown in Figure 2, our method decomposes each generation step into three sub-steps—zig, zag, and generation. It proposes the combination

of two key components: Zig Visual Sharing (ZVS) and Asymmetric Prompt Zigzag Inference (APZI). In the ZVS module, subject-specific visual information that is extracted using a subject-only prompt is injected into the self-attention layers during the generation of each scene. This allows the model to encode subject identity directly into the latent representations without disrupting overall scene composition. To further balance subject consistency with textual fidelity, APZI introduces an asymmetric prompt schedule: the zig and generation steps leverage the one-prompt technique proposed in [29], while the zag step employs a null prompt. This asymmetry prevents overfitting to textual descriptions and provides the model with a dedicated phase to integrate subject information. Together, ZVS and APZI enable effective subject conditioning in the latent space while preserving the model's ability to follow textual prompts, resulting in more coherent and consistent multiscene image generation.

We evaluate our method on two representative text-to-image architectures: the U-Net-based SDXL model [30] and the transformer-based FLUX model [31]. Experimental results demonstrate that our approach significantly improves subject consistency in generated images while maintaining strong alignment with textual prompts, outperforming existing baselines. In summary, our main contributions are:

- We propose a novel asymmetric zigzag sampling algorithm that injects visual semantics during the zig step to enforce subject consistency, uses a null prompt in the zag step to refine the latent space, and applies only text guidance in the generation step to preserve narrative coherence.
- We conduct extensive experiments and benchmark our approach against a range of existing methods to demonstrate its superiority.
- We validate the generalizability of our method by applying it to two widely used text-to-image architectures.

2 Related Work

2.1 Diffusion Models

Diffusion models are a class of generative models that have demonstrated remarkable success in synthesizing high-quality images [1, 30, 4, 3], videos [9, 7, 5, 6], and 3D content [32–34]. These models consist of a forward (noising) process and a reverse (denoising) process. In the forward process, a clean image is progressively corrupted by the addition of Gaussian noise over several steps. During the reverse process, a neural network learns to gradually denoise a random Gaussian sample, ultimately reconstructing a coherent image. The multi-step nature of this inference pipeline offers significant flexibility for controllable or guided generation tasks, such as layout-constrained image synthesis [12, 35, 36, 11], motion-controlled video generation [13, 14], and subject-consistent image personalization [19, 18]. Despite their strong performance, diffusion models remain an active area of research, with ongoing efforts to expand their generative capabilities and improve efficiency. For instance, recent works [37, 38] have proposed accelerated sampling strategies that reduce the number of required inference steps to fewer than 50. Other studies address issues such as exposure bias [39, 40], leading to more robust and higher-fidelity outputs. In general, diffusion models continue to evolve rapidly, and innovations in architecture, training strategies, and sampling techniques push the boundaries of generative modeling.

2.2 Training Based Consistent Text to Image Generation

Since the release of powerful open-source text-to-image diffusion models, generating images featuring consistent subjects — whether for image personalization or visual storytelling — has become an increasingly active area of research. Early approaches [18, 19, 41, 23, 20–22] primarily rely on test-time tuning techniques. These methods typically fine-tune either the entire diffusion network or specific components, learning a unique subject embedding or a small auxiliary network from a few reference images. Although these methods achieve high-quality results in subject-driven image generation, their reliance on test-time optimization poses significant scalability challenges. To address this limitation, a growing body of work [42, 25, 43–47] has explored encoder-based strategies. These methods utilize image encoders to extract subject-specific features from reference images and inject

the resulting embeddings into the diffusion model to guide subject-consistent generation. By avoiding per-subject fine-tuning, such approaches improve efficiency and scalability in real-world applications.

2.3 Training-Free Consistent Text to Image Generation

In parallel with training-based approaches, training-free methods for consistent text-to-image generation have also received significant attention in the visual generation community. A popular direction involves leveraging the attention mechanism to identify subject-relevant visual features, which are then reused to guide subsequent image generation. Initially, such techniques were explored in tasks like appearance transfer [48], image editing [49, 50], and image translation [51]. More recently, ConsisStory [28] adapted this idea to visual storytelling. Specifically, it uses cross-attention scores between text and image tokens to identify subject-relevant visual features in one image, and then injects these features into the generation of subsequent images to maintain subject consistency. In contrast to attention-based guidance, the current state-of-the-art method, 1Prompt1Story [29], maintains subject consistency by composing a unified prompt that combines all scene descriptions. During generation, the model adjusts the weighting of each sub-prompt depending on the scene. Our approach differs from both strategies by introducing **asymmetric guidance** within the zigzag sampling framework. This asymmetric design strengthens subject consistency across generated images while preserving the model's ability to follow textual prompts faithfully.

3 Method

3.1 Preliminary

Latent Diffusion Model. The Latent Diffusion Model (LDM) comprises an autoencoder—consisting of an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$—that maps between pixel space and latent space. A diffusion model ϵ_θ, parameterized by θ, is trained to model the noise in the latent space. For text-conditioned generation, a frozen text encoder τ_ζ is used to embed the input text $\mathcal{P}$ into a dense representation [4]. The diffusion model is trained using the following loss function:

$$L_{LDM} = \mathbb{E}_{c\sim\epsilon(x),\epsilon\sim\mathcal{N}(0,1),t\sim\text{Uniform}(1,T)}[||\epsilon - \epsilon_\theta(z_t, t, \tau_\zeta(\mathcal{P}))||_2^2] \tag{1}$$

During training, the diffusion process involves predicting noise at a randomly sampled time step t, drawn from a uniform distribution. To effectively model this denoising process—especially in the context of text-to-image generation—attention mechanisms play a central role in the architecture of diffusion models. In UNet-based models such as SDXL [30], cross-attention layers facilitate the integration of textual information by aligning latent visual features with text embeddings. This allows the model to generate images that are semantically consistent with the input text. Simultaneously, self-attention layers help capture spatial and semantic relationships within the visual latent space, enabling more coherent and detailed image synthesis. In contrast, the transformer-based FLUX model [31] adopts a different strategy. It uses a unified self-attention mechanism with modality-specific projection layers for visual and textual inputs. This design allows FLUX to fuse semantic information across modalities without relying on explicit cross-attention, leveraging the transformer's strength in modeling complex dependencies. Our method is seamlessly integrated in the UNet-based model SDXL and the transformer-based model FLUX.

3.2 Asymmetry Zigzag Sampling.

To address the challenge of balancing subject consistency and text fidelity in story-based image generation, we propose a novel **Asymmetric Zigzag Sampling** strategy that leverages the strengths of both diffusion-based generation and semantic conditioning. Prior work has introduced zigzag sampling—a method that decomposes each diffusion denoising step into three sub-steps: zig, zag, and generation—to improve the performance of generative models. Inspired by the distinct functional roles identified in recent work [52], where the zig step facilitates exploration, the zag step enables refinement, and the generation step produces the final output, we introduce asymmetry into this process to more precisely regulate the flow of semantic information across steps. In our design, **identity visual semantics** are injected exclusively during the zig (exploration) step to establish strong identity grounding early in the denoising trajectory. The zag step then adjusts the latent space without further visual input, helping to **avoid overfitting to the identity**. Finally, the generation step focuses

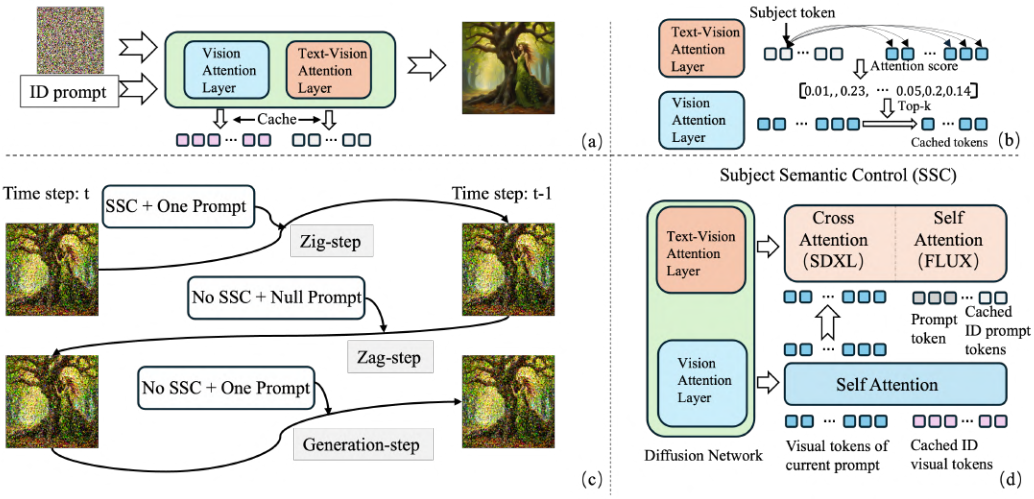

Figure 3: Overview of our proposed pipeline. (a) Identity-guided diffusion inference, where identity prompts are used to cache identity-related visual tokens. (b) Visual token selection module, which leverages attention scores to identify the most relevant tokens for the subject. (c) Illustration of the asymmetric design applied to zigzag sampling. (d) Integration of identity-aware visual information during the zigzag sampling process.

solely on **prompt alignment**, free from additional visual conditioning. This asymmetric configuration allows our method to achieve a superior balance between subject consistency and prompt fidelity, particularly in complex, multi-image story generation tasks. An overview of the proposed method is illustrated in Figure 3, with detailed step-by-step operations described below. Formally, let x_t denote the noisy latent at timestep t, and x_{t-1} the denoised latent. While the zig and generation steps follow the standard forward denoising trajectory, the zag step performs an inverse denoising operation, mapping x_t to a prior latent x_{t+1}. The computation for each sub-step is governed by the noise schedule α_t predefined by the underlying diffusion model.

$$x_{t-1} = \sqrt{\alpha_{t-1}} \frac{x_t - \sqrt{1-\alpha_t}\epsilon_\theta(x_t)}{\sqrt{\alpha_t}} + \sqrt{1-\alpha_{t-1}}\epsilon_\theta(x_t) \qquad \hat{x}_t = \sqrt{\frac{\alpha_t}{\alpha_{t-1}}} x_{t-1} + \sqrt{\alpha_t}\left(\sqrt{\frac{1}{\alpha_t} - 1} - \sqrt{\frac{1}{\alpha_{t-1}} - 1}\right)\epsilon_\theta(x_{t-1})$$

Diffusion denoising step — Inverse diffusion denoising step

Building on the bidirectional nature of zigzag sampling, which enables more effective semantic control during generation, we propose asymmetric zigzag sampling. In our approach, we inject strong subject-specific semantic cues into the intermediate network representation during the zig step, then propagate this information into the noisy latent space through the zag step, while keeping the generation step unchanged. This asymmetric design allows to enhance subject consistency across generated images without compromising the model's ability to accurately follow the textual prompt. We describe our asymmetric setup for zigzag sampling below, and summarize the proposed method in Algorithms 1 and 2 in the appendix.

Asymmetry Visual Guidance. As described above, we inject strong subject-specific cues into the latent representation during the *zig* step. To extract these cues, we first run the diffusion denoising process using an identity-focused prompt that describes only the visual characteristics of the subject. Following prior work [28], we compute text-image attention scores to identify subject-relevant visual tokens at each layer and timestep. These tokens serve as the basis for semantic injection, helping the model preserve subject identity across generated images. We denote the subject-related attention scores in the text-image attention layer as:

$$\mathcal{M}_{\text{subject}} = [s_1^{l,m}, s_2^{l,m}, \cdots, s_n^{l,m}],$$

where l and m denote the layer and timestep indices, respectively. Based on these scores, we cache the top-k key and value projections of visual tokens from the image self-attention layers. This caching process is illustrated in Figure 3 (a) and the selected tokens are defined as:

$$I_{\text{key}}^{l,m} = [i_{\text{key},1}^{l,m}, i_{\text{key},2}^{l,m}, \cdots, i_{\text{key},k}^{l,m}], \quad I_{\text{value}}^{l,m} = [i_{\text{value},1}^{l,m}, i_{\text{value},2}^{l,m}, \cdots, i_{\text{value},k}^{l,m}].$$

During image generation for each prompt, these cached subject tokens are concatenated with the current visual tokens in the image self-attention layers at each layer and timestep. This integration is formalized as follows:

$$K_+^{i,l} = \text{Concatenate}(I_{\text{key}}^{l,m}, K^{i,l}) \qquad V_+^{i,l} = \text{Concatenate}(I_{\text{value}}^{l,m}, V^{i,l})$$

The query tokens remain unchanged. The updated key $K_+^{i,l}$, value $V_+^{i,l}$, and original query tokens are then used to perform standard self-attention, which updates the visual representation in the image attention layer. To avoid compromising the model's ability to follow text prompts, this semantic injection is applied only during the *zig* step. The *zag* and *generation* steps remain unchanged. This process is further illustrated in the Subject Semantic Control (SSC) module shown in Figure 3.

Asymmetry Prompt Guidance. For text-based guidance, we have first adapted an approach proposed in [29], which concatenates all scene-level prompts into a single sentence and dynamically reweights the contribution of each sub-sentence during the generation of the corresponding image. This allows the model to maintain narrative coherence across scenes while emphasizing the relevant textual content at each generation step. In addition, we have integrated the Identity-Preserving Contrastive Alignment (IPCA) technique [29] to further strengthen subject representation through the text prompt, reinforcing the model's ability to retain subject identity throughout the story. However, directly applying these text-guidance strategies within the standard zigzag sampling framework has led to suboptimal results. We hypothesize that the strong text-driven supervision during the *zag* step conflicts with the subject-specific information introduced through visual token injection, potentially overriding or distorting it. This interference degrades subject consistency and overall image quality. To address this issue, we introduce a novel **asymmetric zigzag prompt guidance** strategy. Specifically, we apply the enhanced text-guidance mechanisms only during the *zig* and *generation* steps, where they align well with semantic injection and image refinement. During the *zag* step, we instead use a null prompt to prevent conflicting and allow the subject-specific visual information to propagate unimpeded. This asymmetry helps preserve both textual relevance and subject consistency across the generated sequence.

4 Experiments

Comparison Methods. We integrate our method in two backbone architectures: the UNet-based SDXL [30] and the transformer-based FLUX [31] models. For the SDXL backbone, we compare our approach against several state-of-the-art baselines, including both training-based methods—The Chosen One [53], PhotoMaker [43], and IP-Adapter [24]—and training-free methods—ConsiStory [28], StoryDiffusion [27], and 1Prompt1Story [29]. For the FLUX model, as no prior methods have been demonstrated to support this architecture, we re-implement the most recent and competitive training-free method, 1Prompt1Story, on top of FLUX. We then use this as a baseline for comparison against our method. This dual-platform evaluation demonstrates the adaptability and effectiveness of our approach across diverse diffusion architectures.

Evaluation Metrics. Following prior work [29, 28, 27], we evaluate all models along two key dimensions: **prompt alignment** and **subject consistency**. To assess prompt alignment, we use the CLIP image and text encoders to compute the average CLIP-Score [54] between each generated image and its corresponding prompt. This metric reflects how well the visual content of the generated image aligns with the intended textual description. For subject consistency, we adopt two complementary metrics from previous studies [29]: DreamSim [55] and CLIP-I [54]. DreamSim measures perceptual similarity and has shown strong correlation with human judgment in evaluating visual coherence. CLIP-I computes the average cosine similarity between CLIP image embeddings, capturing the consistency of subject identity across different images. To ensure that subject consistency is not influenced by variations in background content, we follow prior evaluation protocols and apply CarveKit [56] to remove the background from each generated image. The removed regions are then filled with random noise, isolating the subject and allowing for a more focused and accurate evaluation.

5 Results

Quantitative Comparison. Table 1 presents a quantitative comparison between our proposed method and previously discussed approaches. As shown, our Asymmetry ZigZag Sampling technique achieves the best overall performance across all evaluation metrics among training-free visual storytelling methods. When compared with training-based methods, our approach outperforms both PhotoMaker and The Chosen One across all evaluation metrics. While the IP-Adapter method demonstrates the highest performance in DreamSim and CLIP-I scores, our method achieves a comparable CLIP-I score and significantly narrows the gap in DreamSim performance between training-free and training-based methods reducing it by 63.27%. Although the IP-Adapter excels in subject consistency metrics (CLIP-I and DreamSim), it performs considerably worse in text-alignment metrics such as CLIP-T. This discrepancy may be due to the IP-Adapter's tendency to overemphasize the subject, often generating images with highly similar or repetitive layouts, as illustrated in the figure 4.

Method	Base Model	Train-Free	CLIP-T↑	CLIP-I↑	DreamSim↓	Steps
Vanilla SDXL	-	-	0.9074	0.8165	0.5292	50
Vanilla FLUX	-	-	0.8977	0.8494	0.3888	28
The Chosen One	SDXL	✗	0.7614	0.7831	0.4929	35
PhotoMaker	SDXL	✗	0.8651	0.8465	0.3996	50
IP-Adapter	SDXL	✗	0.8458	**0.9429**	**0.1462**	30
ConsiStory	SDXL	✓	0.8769	0.8737	0.3188	50
StoryDiffusion	SDXL	✓	0.8877	0.8755	0.3212	50
1Prompt1Story	SDXL	✓	0.8942	0.9117	0.1993	50
Ours	SDXL	✓	**0.8946**	0.923	0.1798	50
1Prompt1Story	FLUX	✓	0.8716	0.9118	0.1957	28
Ours	FLUX	✓	**0.8949**	**0.9216**	**0.1843**	28

Table 1: Quantitative Comparison. We report quantitative results for different methods. For the SDXL architecture, the best-performing value is highlighted in bold, while the second-best is indicated with a box. For the FLUX model, only the best result is highlighted in bold. The baseline models—vanilla SDXL and vanilla FLUX—are included as references but are excluded from the comparative ranking.

User Study. While automatic evaluation metrics offer a useful quantitative assessment, they can be biased due to their reliance on pretrained models. To further validate the effectiveness of our proposed method, we conducted a human evaluation study. We randomly selected 30 prompts from the benchmark dataset and generated corresponding image sequences using all competing methods. Twenty participants were invited for the user study. For each participant, a custom program randomly selected 20 out of the 30 prompts, and presented four resulting image sequences obtained with different methods for each selected prompt. Participants were asked to choose their preferred image sequence based on three criteria: identity consistency, prompt alignment, and overall image quality. As shown in Table 2, images generated by our method received the highest preference from participants, indicating strong alignment with human judgment. Further details of the user study protocol and interface can be found in the Appendix.

Method	**IP-Adapter**	**ConsiStory**	**StoryDiffusion**	**1Pormpt1Story**	**Ours**
Percent (%)↑	5.15	19.18	16.25	24.50	**35.02**

Table 2: Human preference comparison among different methods.

Qualitative Comparison. Figures 4 and 6 present qualitative comparisons of our method against existing approaches. For the SDXL model, we compare our method with IP-Adapter [24], ConsisTory [28], StoryDiffusion [27], and 1Prompt1Story [29]. For the FLUX model, we evaluate our method against the training-free, state-of-the-art 1Prompt1Story approach. As shown in Figure 4, IP-Adapter tends to overemphasize identity consistency at the expense of faithfully representing the input text. Meanwhile, ConsisTory, StoryDiffusion, and 1Prompt1Story struggle to maintain subject

consistency across different images. In contrast, our method achieves a well-balanced performance, effectively preserving subject identity while accurately following the text descriptions. Figure 6 further underscores the superiority of our approach over the 1Prompt1Story model in terms of both subject consistency and text alignment. Additionally, it highlights the generalizability of our method, which maintains strong performance across diverse network architectures.

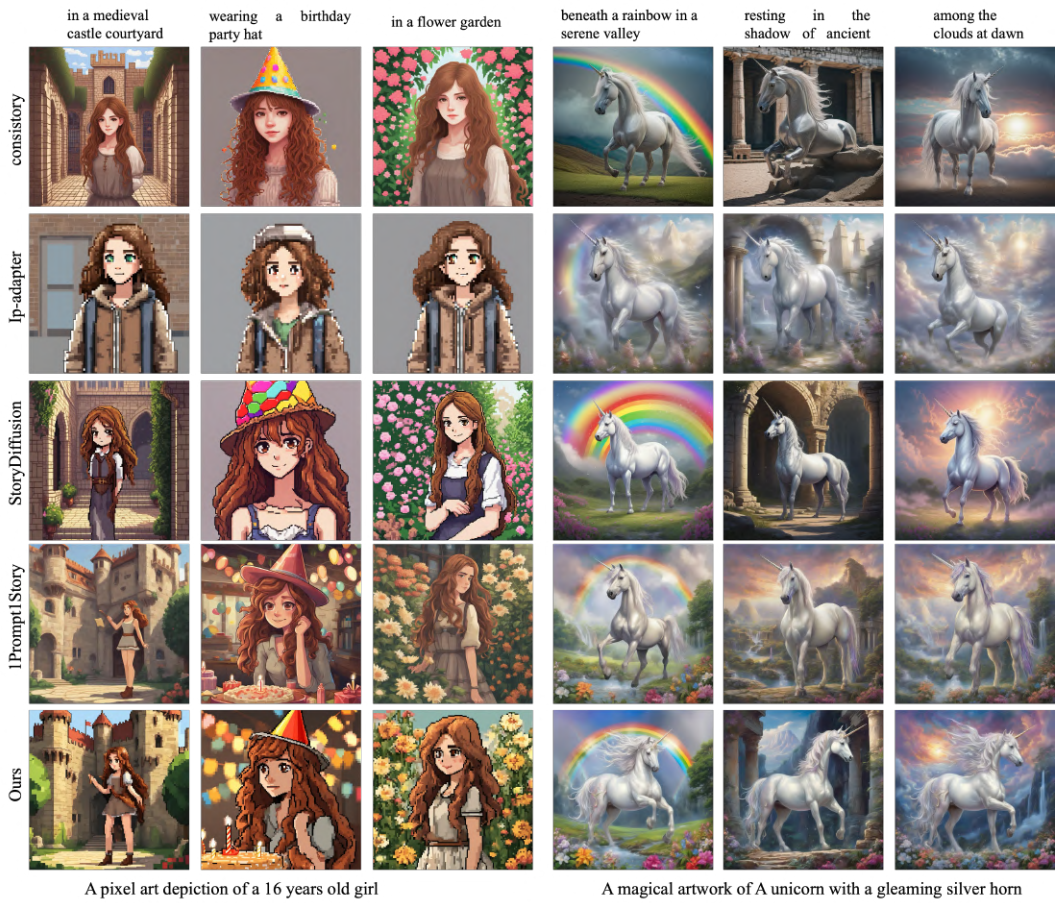

Figure 4: Qualitative Results Using the SDXL Backbone. We compare our method with four baselines: ConsisStory, IP-Adapter, StoryDiffusion, and 1Prompt1Story. The identity prompt is shown at the bottom, while individual image prompts are displayed above each corresponding image. Our method demonstrates a strong balance between maintaining subject consistency and adhering to textual prompts. In contrast, the baseline methods often struggle—either failing to preserve the subject's identity or deviating from the given text descriptions.

Method	CLIP-T↑	CLIP-I↑	DreamSim↓
Asymmetry	0.8946	0.923	0.1798
Zig-gen	0.8534	0.919	0.1801
Zig-zag	0.8944	0.8916	0.2121
All	0.8788	0.8833	0.2312

Table 3: Quantitative comparisons of different zigzag sampling designs.

Figure 5: Qualitative comparisons of different zigzag sampling designs.

Ablation study:Effect of Asymmetric Visual Injection. We conduct experiments to evaluate the effectiveness of our proposed asymmetric visual injection design by comparing it with three alternative visual injection strategies: (1) zig-gen symmetric sampling, where visual semantics are injected during

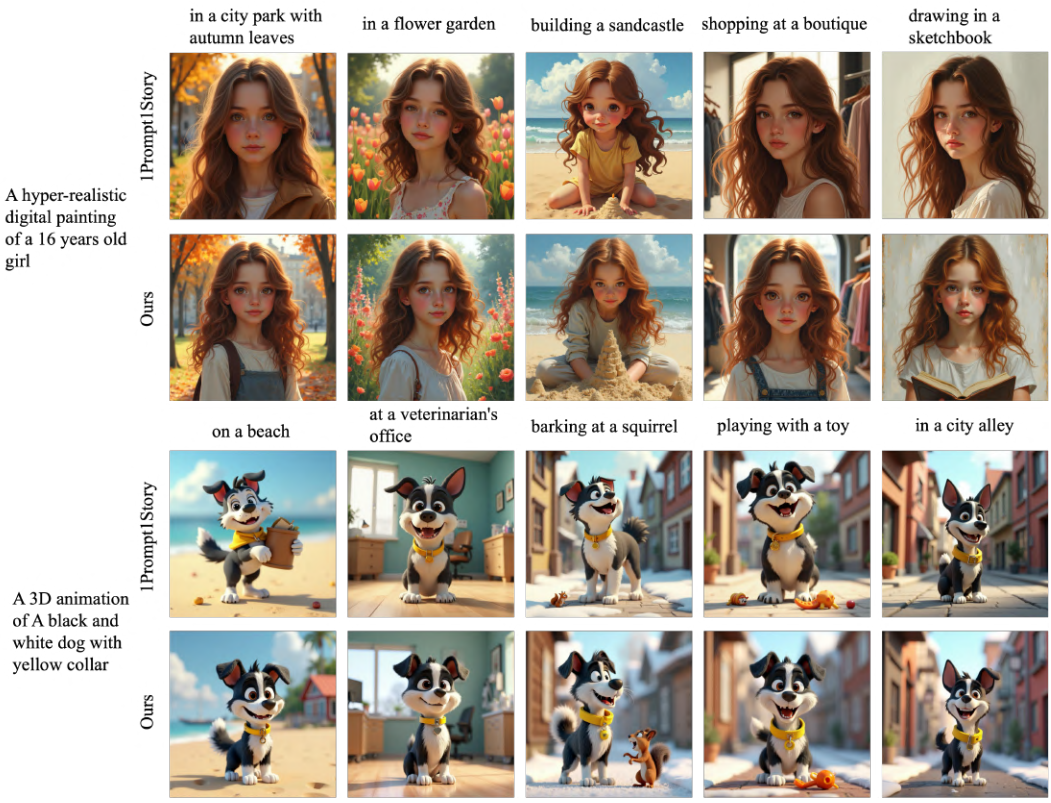

Figure 6: Qualitative Results: We compare our method with 1Prompt1Story using the FLUX backbone. Across varying prompts, our method demonstrates superior ability to preserve subject identity, highlighting its robustness in maintaining consistency.

the zig and generation steps but omitted in the zag step; (2) zig-zag symmetric sampling, which injects visual information in the zig and zag steps while excluding it during generation; and (3) fully symmetric sampling, where visual injection is applied across all three steps—zig, zag, and generation. These comparisons allow us to isolate the impact of asymmetric conditioning on balancing subject identity consistency and text-image alignment. Table 3 shows that fully symmetric sampling performs worst, likely due to over-conditioning, which introduces visual artifacts also observable in Figure 5. While the zig-gen strategy achieves higher identity similarity than zig-zag—likely due to visual injection during generation—it suffers from poor text alignment. Conversely, zig-zag provides better alignment at the expense of identity consistency. Our asymmetric visual conditioning design achieves the best overall balance between these two competing objectives, i.e., identity preservation and accurate adherence to textual prompts. These results validate the effectiveness of the proposed asymmetric strategy.

Ablation Study: Effect of Asymmetric Prompt Conditioning. In our method, we employ prompt conditioning in the zig and generation steps, while using a null (empty) prompt in the zag step. This asymmetric design is motivated by the intuition that removing the prompt in the zag step helps prevent overfitting to the text condition and allows the model to better retain the subject-specific information introduced during the zig step. To validate the effectiveness of this design, we compare a symmetric variant—where prompt conditioning is applied uniformly across all steps—with our asymmetric configuration. As shown in Table 4, the asymmetric design consistently achieves superior performance, with CLIP-T and CLIP-I scores improving from $0.9098/0.8841$ to $0.923/0.8946$, and DreamSim (lower is better) decreasing from 0.2251 to 0.1789. These results confirm that the asymmetric prompt conditioning facilitates stronger

	CLIP-T↑	CLIP-I↑	DreamSim↓
Symmetry	0.9098	0.8841	0.2251
Asymmetry	0.923	0.8946	0.1798

Table 4: Quantitative comparisons of symmetry (using prompt condition in zag step) and asymmetry (using null prompt conditionin zag step) design.

semantic alignment with the textual prompt while preserving visual fidelity and preventing excessive prompt dependency, making it a more effective and balanced design choice for our framework.

Ablation Study: The Influence of Hyperparameter k In our visual injection module, we select the top-k visual tokens based on attention scores for injection during the sampling process. We evaluate the impact of varying k values (ranging from 0.2 to 0.8) on generation performance (Table 5). The results show that increasing k generally improves image similarity but slightly reduces text alignment. Notably, a value of $k = 0.2$ achieves the best balance between these two objectives. Although different k values affect both metrics, the variations are minor, suggesting that our method is robust to the choice of k.

k	CLIP-T↑	CLIP-I↑	DreamSim↓
0.2	0.8946	0.923	0.1798
0.4	0.8931	0.9245	0.1799
0.6	0.8921	0.9258	0.1801
0.8	0.8924	0.9287	0.1827

Table 5: The influence of hyperparameter k on the generation performance.

Inference Time Comparison. Our method can also be regarded as a form of test-time scaling in diffusion models. Compared with the standard diffusion inference process, our approach decomposes each inference step into three sub-steps—zig, zag, and generation—which inevitably introduces additional computational overhead and leads to longer inference time. To examine this effect, we evaluated the inference speed of our method against the regular diffusion process using both a UNet-based SDXL model and a transformer-based FLUX model. For SDXL with 50 sampling steps, the baseline model required 7.44 seconds per image, while our method took 21.33 seconds. Similarly, for FLUX with 28 sampling steps, the baseline achieved 1.50 seconds per image, whereas our method required 5.50 seconds. These results indicate that our approach increases the inference time by approximately 2.9 on SDXL and 3.7 on FLUX. Despite the additional computational cost, the improved subject consistency across images in visual storytelling justifies the longer inference time as a reasonable trade-off for enhanced generative performance.

Other Applications. Similar to the One-Prompt-One-Story approach [29], our method can be naturally extended to generate long stories containing an arbitrary number of images. This is achieved by applying a sliding-window strategy over the sequence of prompts, where each window of prompts is processed using our proposed method to ensure coherent subject and style consistency across adjacent images. By iteratively moving the window through the entire prompt sequence, our approach can maintain both local and global narrative consistency, enabling the generation of super-long visual stories. A visualization of such an extended story is provided in the appendix, demonstrating the capability of our method to handle narratives of considerable length while preserving high-quality and consistent image generation.

6 Conclusion

We propose an asymmetric zigzag sampling strategy to address the challenge of training-free visual storytelling, achieving a strong balance between subject consistency and text alignment. By leveraging the distinct roles of zig (exploration), zag (adjustment), and generation, our method injects visual semantics only where needed, enabling coherent, identity-consistent outputs without retraining. This approach also highlights the potential of sampling strategies as a promising direction for future research in controllable image generation.

7 Limitations

While our proposed method demonstrates strong performance in maintaining subject consistency and adhering to textual descriptions in visual story generation, it is not without limitations. First, the use of zigzag sampling—comprising three sub-steps per generation step—introduces additional computational overhead, which may increase inference time compared to standard sampling strategies. Second, although our approach is compatible with both the SDXL and FLUX architectures, our experiments indicate that it integrates more seamlessly with the FLUX model. In some cases, we observe occasional failures or reduced performance when applied to SDXL, suggesting that further architecture-specific optimization could enhance robustness and generalizability. Nonetheless, these limitations do not detract from the overall effectiveness of our method and highlight promising directions for future improvement.

A Acknowledgement

This work is funded by the CALCULUS project (European Research Council Advanced Grant H2020-ERC-2017-ADG 788506) and the Flanders AI Research Program.

B Boarder Impacts

Visual generative techniques, particularly text-to-image (T2I) models, hold significant potential for producing coherent visual content across diverse scenarios, making them highly applicable to downstream tasks such as storytelling and personalized content creation. One of the most challenging aspects in this domain is the consistent synthesis of characters across varying contexts—a problem that existing methods continue to struggle with, as discussed in this paper. Our proposed approach addresses this challenge by effectively balancing subject consistency and prompt fidelity, allowing users to generate coherent story sequences featuring the same character while closely adhering to their provided descriptions. In addition, our exploration of the unique structure of zigzag sampling introduces a new perspective on its utility in diffusion-based generation, offering valuable insights that may inspire future research into more controllable and semantically aware generative models.

The application of text-to-image models in visual storytelling, while creatively empowering, introduces significant ethical, privacy, and security risks. A major concern is the non-consensual creation of fictional narratives featuring real individuals, where realistic and consistent character generation enables the fabrication of defamatory, misleading, or harmful visual stories—such as fake memoirs, satirical comics, or illustrated scenarios depicting private citizens or public figures in compromising situations—without their knowledge or consent. The very capability to maintain character coherence across scenes, which enhances narrative immersion, can be exploited to produce persuasive, long-form synthetic content that blurs the line between fiction and reality, facilitating disinformation campaigns or emotional manipulation. Furthermore, privacy is compromised when models trained on unconsented web-scraped data generate characters closely resembling real people, effectively creating digital doppelgängers embedded in fictional universes. These capabilities also pose security threats, as coherent AI-generated visual stories can be weaponized for influence operations, identity exploitation, or viral misinformation, undermining trust and personal autonomy. Without robust safeguards—such as provenance tracking, consent filters, and transparent content policies—AI-driven visual storytelling risks enabling large-scale narrative abuse with profound societal consequences.

C Usage of LLMS

We only use large language models (LLMs) for writing assistance, such as correcting grammar, fixing typos, and improving clarity.

D Implementation Details

We implement our method on two open-source models: SDXL and FLUX. For SDXL, we use the *stabilityai/stable-diffusion-xl-base-1.0* version, and for FLUX, we adopt the *black-forest-labs/FLUX.1-dev* version. All baseline methods—including IP-Adapter [24], Consistory [28], StoryDiffusion [27], and 1Prompt1Story [29]—are reproduced using their official open-source implementations with default hyperparameters. Since IP-Adapter is designed for image-conditional generation, we adapt it to our setting by first generating an identity image using SDXL with the given identity prompt. This generated image is then used as the conditioning input for IP-Adapter to produce images guided by different prompts.

For the implementation of our method on the SDXL model, we cache visual tokens only from the mid and upper layers across all steps. Accordingly, feature injection during the zig step is also limited to these layers. We use a classifier-free guidance scale of 5.5 for both the zig and generation steps, and set it to 0 during the zag step. All experiments are conducted on a single NVIDIA A100 GPU.

For the FLUX model, which differs architecturally from SDXL by separating text-image and image-image interaction stages, we adopt a different strategy. FLUX begins with several layers of text-image interaction, followed by layers of purely image-based interaction. To cache visual tokens, we first average the attention scores from the early text-image interaction layers to identify subject-relevant

visual features. These selected tokens are then used for feature injection across all image-image interaction layers. Experiments for FLUX are also run on a single NVIDIA A100 GPU.

Algorithm 1 Identity Visual Token Cache (Subject token extraction & top-k selection)

Require: identity prompt P_{id}, pretrained diffusion model ϵ_θ, text encoder τ_ζ, time steps $\{m\}$ (used for identity extraction), layers $\mathcal{L}$, top-k ratio k

Ensure: cached key tokens $\mathcal{I}^{\text{key}} = \{I^{\ell,m}_{\text{key}}\}$ and value tokens $\mathcal{I}^{\text{value}} = \{I^{\ell,m}_{\text{value}}\}$

Compute text embedding $T_{\text{id}} \leftarrow \tau_\zeta(P_{\text{id}})$.
for each layer $\ell \in \mathcal{L}$ and timestep m used for identity extraction **do**
  Run denoising pass (or inference pass) with prompt P_{id} to obtain intermediate visual tokens at (ℓ, m).
  Compute text-image attention scores $S^{\ell,m} = \text{AttentionScores}(\text{visual tokens}, T_{\text{id}})$.
  Identify top-k indices by score: $J^{\ell,m} \leftarrow \text{TopKIndices}(S^{\ell,m}, k)$.
  Extract corresponding key / value projections:

$$I^{\ell,m}_{\text{key}} \leftarrow \{i^{\ell,m}_{\text{key},j} : j \in J^{\ell,m}\}, \quad I^{\ell,m}_{\text{value}} \leftarrow \{i^{\ell,m}_{\text{value},j} : j \in J^{\ell,m}\}.$$

  Store $I^{\ell,m}_{\text{key}}$ into $\mathcal{I}^{\text{key}}$ and $I^{\ell,m}_{\text{value}}$ into $\mathcal{I}^{\text{value}}$.
end for
return $\mathcal{I}^{\text{key}}, \mathcal{I}^{\text{value}}$

Algorithm 2 Asymmetric Zigzag Sampling with Zig Visual Sharing (ZVS) & Asymmetric Prompt Zigzag Inference (APZI)

Require: target prompt P, identity token caches $\mathcal{I}^{\text{key}}, \mathcal{I}^{\text{value}}$ (from Alg. 1), diffusion model ϵ_θ, text encoder τ_ζ, full zigzag time schedule $t = T, \ldots, 1$, and its noise coefficients α_t

Ensure: latent x_0 (decoded to image by decoder D)

Compute full prompt embedding $T \leftarrow \tau_\zeta(P)$ (can use one-prompt fusion / reweighting as in paper).
Initialize noisy latent $x_T \sim \mathcal{N}(0, I)$.
for each diffusion step index $t = T, T-1, \ldots, 1$ **do**
  Zig step (forward denoise + visual injection):
    1. Compute standard denoising prediction $\hat{\epsilon} \leftarrow \epsilon_\theta(x_t, t, T)$ using prompt embedding T.
    2. Perform denoising update (forward) to get intermediate latent x_{t-1}
    3. **Inject identity visual tokens** into self-attention layers: for each layer ℓ used,

$$K^{\ell,+} \leftarrow \text{Concat}\big(I^{\ell,m}_{\text{key}},\ K^\ell\big), \quad V^{\ell,+} \leftarrow \text{Concat}\big(I^{\ell,m}_{\text{value}},\ V^\ell\big),$$

    where K^ℓ, V^ℓ are the current layer key/value projections and $I^{\ell,m}_{\cdot}$ come from the caches. Keep queries unchanged. (Apply only in Zig.)
  Zag step (inverse denoise — null prompt / no text guidance):
    1. Use null prompt.
    2. Perform inverse denoising mapping to propagate the injected identity into the noisy latent:

$$\tilde{x}_t \leftarrow \text{InverseDenoise}(x_{t-1},\ t-1 \to t,\ \epsilon_\theta,\ \text{null prompt}),$$

    3. (No visual injection in Zag.)
  Generation step (final denoise with text guidance):
    1. Use full prompt embedding T to compute ϵ-prediction on $\tilde{x}_t$.
    2. Perform forward denoising to obtain x^{final}_{t-1} (standard step).
    3. Set $x_{t-1} \leftarrow x^{\text{final}}_{t-1}$ and continue.
end for
Decode x_0 to image: $I \leftarrow D(x_0)$ and return I.

E More Details about User Study

We conducted a user study to evaluate our method in comparison with four existing approaches: IP-Adapter [24], Consistory Model [28], Story Diffusion [27], and 1Prompt1Story [29]. All models were used to generate images based on prompts from the *ConsiStory+* benchmark, using the same random seeds as reported in their respective papers to ensure a fair comparison.

From the generated dataset, we randomly selected 30 prompts, each associated with 4 images. For each participant in the user study, the system randomly sampled 20 out of these 30 prompts to form an evaluation set. Before starting the evaluation, users were briefed on three key criteria:

- **Identity Consistency**: Measures whether the same character or subject appears consistently across all images for a given prompt.
- **Prompt Alignment**: Assesses how well each image reflects the content and intent of the original text prompt.
- **Image Quality**: Evaluates the overall visual quality, including clarity, detail, and aesthetic appeal.

To minimize potential bias, the presentation order of the five methods was randomized for each question in the study interface. Figure 10 presented the user interface of our user study system.

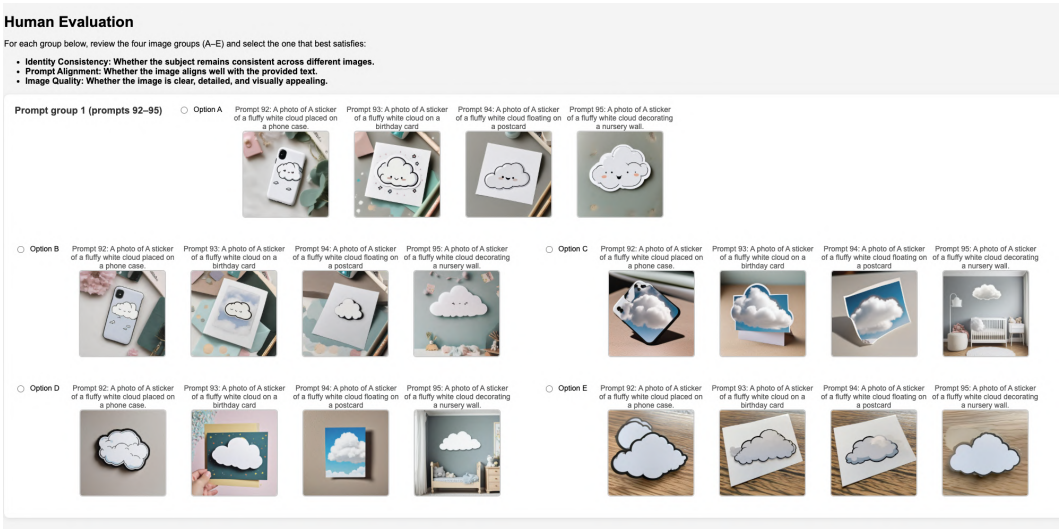

Figure 7: A visualization of the user interface used in our user study system.

F More Visualizations

Figure 8: More visualzations of results generated using FLUX model.

Figure 9: More visualzations of results generated using SDXL model.

G Long Story Visualizations

Figure 10: Long story generated using FLUX model

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

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

   Answer: [Yes]

   Justification: The main claims in the abstract and introduction are aligned with the key contributions of the paper.

   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: Due to the limited space, we discuss the limitation in the appendix.

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: [NA]

   Justification: There is no theoretical result in the paper.

   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: Due to the limited space, we explain all our implementation details in the appendix.

   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: [NA]

   Justification: We will open-source our code and release detailed instructions upon acceptance of the paper.

   - 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: Due to the limited space, we explains all the details in the appendix.

   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: [NA]

   Justification: The computational cost of each experiment is substantial, and to the best of our knowledge, prior work in this field has not reported statistical significance testing. In line with these works, we do not include statistical significance analysis in our results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: We explains all these details in the appendix.

   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: We are sure that our research in this paper conform with the NeurIPS code of ethics.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: We discuss these impacts in the appendix.

    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 use open-source pretrained model.

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: The pretraiend model we used are public available.

Guidelines:

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

13. **New assets**

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

Answer: [No]

Justification: We do not introduce new assets in this work.

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 provide this information in the appendix.

Guidelines:

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

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

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

Answer: [NA]

Justification: there is not such issue in this work.

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.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: We describe the usage of LLMs in the appendix.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

References

[1] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel *et al.*, "Scaling rectified flow transformers for high-resolution image synthesis," *URL https://arxiv. org/abs/2403.03206*, vol. 2, 2024.

[2] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, p. 3, 2022.

[3] J. Chen, J. Yu, C. Ge, L. Yao, E. Xie, Y. Wu, Z. Wang, J. Kwok, P. Luo, H. Lu *et al.*, "Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis," *arXiv preprint arXiv:2310.00426*, 2023.

[4] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[5] W. Kong, Q. Tian, Z. Zhang, R. Min, Z. Dai, J. Zhou, J. Xiong, X. Li, B. Wu, J. Zhang *et al.*, "Hunyuanvideo: A systematic framework for large video generative models," *arXiv preprint arXiv:2412.03603*, 2024.

[6] A. Wang, B. Ai, B. Wen, C. Mao, C.-W. Xie, D. Chen, F. Yu, H. Zhao, J. Yang, J. Zeng *et al.*, "Wan: Open and advanced large-scale video generative models," *arXiv preprint arXiv:2503.20314*, 2025.

[7] G. Ma, H. Huang, K. Yan, L. Chen, N. Duan, S. Yin, C. Wan, R. Ming, X. Song, X. Chen *et al.*, "Step-video-t2v technical report: The practice, challenges, and future of video foundation model," *arXiv preprint arXiv:2502.10248*, 2025.

[8] Sand-AI, "Magi-1: Autoregressive video generation at scale," 2025. [Online]. Available: https://static.magi.world/static/files/MAGI_1.pdf

[9] F. Bao, C. Xiang, G. Yue, G. He, H. Zhu, K. Zheng, M. Zhao, S. Liu, Y. Wang, and J. Zhu, "Vidu: a highly consistent, dynamic and skilled text-to-video generator with diffusion models," *arXiv preprint arXiv:2405.04233*, 2024.

[10] A. Polyak, A. Zohar, A. Brown, A. Tjandra, A. Sinha, A. Lee, A. Vyas, B. Shi, C.-Y. Ma, C.-Y. Chuang, D. Yan, D. Choudhary, D. Wang, G. Sethi, G. Pang, H. Ma, I. Misra, J. Hou, J. Wang, K. Jagadeesh, K. Li, L. Zhang, M. Singh, M. Williamson, M. Le, M. Yu, M. K. Singh, P. Zhang, P. Vajda, Q. Duval, R. Girdhar, R. Sumbaly, S. S. Rambhatla, S. Tsai, S. Azadi, S. Datta, S. Chen, S. Bell, S. Ramaswamy, S. Sheynin, S. Bhattacharya, S. Motwani, T. Xu, T. Li, T. Hou, W.-N. Hsu, X. Yin, X. Dai, Y. Taigman, Y. Luo, Y.-C. Liu, Y.-C. Wu, Y. Zhao, Y. Kirstain, Z. He, Z. He, A. Pumarola, A. Thabet, A. Sanakoyeu, A. Mallya, B. Guo, B. Araya, B. Kerr, C. Wood, C. Liu, C. Peng, D. Vengertsev, E. Schonfeld, E. Blanchard, F. Juefei-Xu, F. Nord, J. Liang, J. Hoffman, J. Kohler, K. Fire, K. Sivakumar, L. Chen, L. Yu, L. Gao, M. Georgopoulos, R. Moritz, S. K. Sampson, S. Li, S. Parmeggiani, S. Fine, T. Fowler, V. Petrovic, and Y. Du, "Movie gen: A cast of media foundation models," 2025. [Online]. Available: https://arxiv.org/abs/2410.13720

[11] Y. Li, H. Liu, Q. Wu, F. Mu, J. Yang, J. Gao, C. Li, and Y. J. Lee, "Gligen: Open-set grounded text-to-image generation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 22 511–22 521.

[12] G. Zheng, X. Zhou, X. Li, Z. Qi, Y. Shan, and X. Li, "Layoutdiffusion: Controllable diffusion model for layout-to-image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 490–22 499.

[13] M. Li, B. Wan, M.-F. Moens, and T. Tuytelaars, "Animate your motion: Turning still images into dynamic videos," in *European Conference on Computer Vision*. Springer, 2024, pp. 409–425.

[14] D. Geng, C. Herrmann, J. Hur, F. Cole, S. Zhang, T. Pfaff, T. Lopez-Guevara, C. Doersch, Y. Aytar, M. Rubinstein *et al.*, "Motion prompting: Controlling video generation with motion trajectories," *arXiv preprint arXiv:2412.02700*, 2024.

[15] J. Sun, M. Li, Z. Chen, Y. Zhang, S. Wang, and M.-F. Moens, "Contrast, attend and diffuse to decode high-resolution images from brain activities," *Advances in Neural Information Processing Systems*, vol. 36, pp. 12 332–12 348, 2023.

[16] J. Sun, M. Li, and M.-F. Moens, "Decoding realistic images from brain activity with contrastive self-supervision and latent diffusion," *arXiv preprint arXiv:2310.00318*, 2023.

[17] ——, "Neuralflix: A simple while effective framework for semantic decoding of videos from non-invasive brain recordings," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 7, 2025, pp. 7096–7104.

[18] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 22 500–22 510.

[19] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-Or, "An image is worth one word: Personalizing text-to-image generation using textual inversion," *arXiv preprint arXiv:2208.01618*, 2022.

[20] N. Kumari, B. Zhang, R. Zhang, E. Shechtman, and J.-Y. Zhu, "Multi-concept customization of text-to-image diffusion," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 1931–1941.

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

[22] Y. Alaluf, E. Richardson, G. Metzer, and D. Cohen-Or, "A neural space-time representation for text-to-image personalization," *ACM Transactions on Graphics (TOG)*, vol. 42, no. 6, pp. 1–10, 2023.

[23] M. Li, T. Qu, T. Tuytelaars, and M.-F. Moens, "Towards more accurate personalized image generation: Addressing overfitting and evaluation bias," *arXiv preprint arXiv:2503.06632*, 2025.

[24] H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang, "Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models," *arXiv preprint arXiv:2308.06721*, 2023.

[25] S. Cui, J. Guo, X. An, J. Deng, Y. Zhao, X. Wei, and Z. Feng, "Idadapter: Learning mixed features for tuning-free personalization of text-to-image models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 950–959.

[26] D. Li, J. Li, and S. Hoi, "Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing," *Advances in Neural Information Processing Systems*, vol. 36, pp. 30 146–30 166, 2023.

[27] Y. Zhou, D. Zhou, M.-M. Cheng, J. Feng, and Q. Hou, "Storydiffusion: Consistent self-attention for long-range image and video generation," *Advances in Neural Information Processing Systems*, vol. 37, pp. 110 315–110 340, 2024.

[28] Y. Tewel, O. Kaduri, R. Gal, Y. Kasten, L. Wolf, G. Chechik, and Y. Atzmon, "Training-free consistent text-to-image generation," *ACM Transactions on Graphics (TOG)*, vol. 43, no. 4, pp. 1–18, 2024.

[29] T. Liu, K. Wang, S. Li, J. van de Weijer, F. S. Khan, S. Yang, Y. Wang, J. Yang, and M.-M. Cheng, "One-prompt-one-story: Free-lunch consistent text-to-image generation using a single prompt," *arXiv preprint arXiv:2501.13554*, 2025.

[30] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, "Sdxl: Improving latent diffusion models for high-resolution image synthesis," *arXiv preprint arXiv:2307.01952*, 2023.

[31] B. F. Labs, "Flux," https://github.com/black-forest-labs/flux, 2024.

[32] C.-H. Lin, J. Gao, L. Tang, T. Takikawa, X. Zeng, X. Huang, K. Kreis, S. Fidler, M.-Y. Liu, and T.-Y. Lin, "Magic3d: High-resolution text-to-3d content creation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 300–309.

[33] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall, "Dreamfusion: Text-to-3d using 2d diffusion," *arXiv preprint arXiv:2209.14988*, 2022.

[34] B. Tang, J. Wang, Z. Wu, and L. Zhang, "Stable score distillation for high-quality 3d generation," *arXiv preprint arXiv:2312.09305*, 2023.

[35] C. Mou, X. Wang, L. Xie, Y. Wu, J. Zhang, Z. Qi, and Y. Shan, "T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 38, no. 5, 2024, pp. 4296–4304.

[36] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 3836–3847.

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

[38] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu, "Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps," *Advances in Neural Information Processing Systems*, vol. 35, pp. 5775–5787, 2022.

[39] M. Li, T. Qu, R. Yao, W. Sun, and M.-F. Moens, "Alleviating exposure bias in diffusion models through sampling with shifted time steps," *arXiv preprint arXiv:2305.15583*, 2023.

[40] M. Ning, M. Li, J. Su, A. A. Salah, and I. O. Ertugrul, "Elucidating the exposure bias in diffusion models," *arXiv preprint arXiv:2308.15321*, 2023.

[41] L. Han, Y. Li, H. Zhang, P. Milanfar, D. Metaxas, and F. Yang, "Svdiff: Compact parameter space for diffusion fine-tuning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 7323–7334.

[42] D. Li, J. Li, and S. Hoi, "Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing," *Advances in Neural Information Processing Systems*, vol. 36, pp. 30 146–30 166, 2023.

[43] Z. Li, M. Cao, X. Wang, Z. Qi, M.-M. Cheng, and Y. Shan, "Photomaker: Customizing realistic human photos via stacked id embedding," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 8640–8650.

[44] Z. Zhou, J. Li, H. Li, N. Chen, and X. Tang, "Storymaker: Towards holistic consistent characters in text-to-image generation," *arXiv preprint arXiv:2409.12576*, 2024.

[45] J. Nam, S. Son, Z. Xu, J. Shi, D. Liu, F. Liu, A. Misraa, S. Kim, and Y. Zhou, "Visual persona: Foundation model for full-body human customization," *arXiv preprint arXiv:2503.15406*, 2025.

[46] Y. Wei, Y. Zhang, Z. Ji, J. Bai, L. Zhang, and W. Zuo, "Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 15 943–15 953.

[47] W. Chen, H. Hu, Y. Li, N. Ruiz, X. Jia, M.-W. Chang, and W. W. Cohen, "Subject-driven text-to-image generation via apprenticeship learning," *Advances in Neural Information Processing Systems*, vol. 36, pp. 30 286–30 305, 2023.

[48] Y. Alaluf, D. Garibi, O. Patashnik, H. Averbuch-Elor, and D. Cohen-Or, "Cross-image attention for zero-shot appearance transfer," in *ACM SIGGRAPH 2024 Conference Papers*, 2024, pp. 1–12.

[49] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or, "Prompt-to-prompt image editing with cross attention control.(2022)," *URL https://arxiv.org/abs/2208.01626*, vol. 1, 2022.

[50] N. Tumanyan, M. Geyer, S. Bagon, and T. Dekel, "Plug-and-play diffusion features for text-driven image-to-image translation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1921–1930.

[51] G. Parmar, K. Kumar Singh, R. Zhang, Y. Li, J. Lu, and J.-Y. Zhu, "Zero-shot image-to-image translation," in *ACM SIGGRAPH 2023 conference proceedings*, 2023, pp. 1–11.

[52] Z. Zhou, S. Shao, L. Bai, Z. Xu, B. Han, and Z. Xie, "Golden noise for diffusion models: A learning framework," *arXiv preprint arXiv:2411.09502*, 2024.

[53] O. Avrahami, A. Hertz, Y. Vinker, M. Arar, S. Fruchter, O. Fried, D. Cohen-Or, and D. Lischinski, "The chosen one: Consistent characters in text-to-image diffusion models," in *ACM SIGGRAPH 2024 conference papers*, 2024, pp. 1–12.

[54] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi, "Clipscore: A reference-free evaluation metric for image captioning," *arXiv preprint arXiv:2104.08718*, 2021.

[55] S. Fu, N. Tamir, S. Sundaram, L. Chai, R. Zhang, T. Dekel, and P. Isola, "Dreamsim: Learning new dimensions of human visual similarity using synthetic data," *arXiv preprint arXiv:2306.09344*, 2023.

[56] N. Selin, "Carvekit: Automated high-quality background removal framework," 2023.