SELECTIVE ATTENTION FOR TRAINING-FREE CONSISTENT TEXT-TO-IMAGE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Generating images with consistent object appearances across different scenes remains a fundamental challenge in text-to-image diffusion models. Existing approaches either require expensive fine-tuning for each subject or suffer from feature leakage when unconditionally sharing attention across images, where irrelevant backgrounds contaminate the target object's appearance. Unconstrained cross-image feature aggregation can introduce irrelevant background information, while selective connections can mitigate this issue. Motivated by this, we propose **SelectDiff**, a training-free framework based on similarity-guided selective attention. Instead of fusing all cross-image tokens, we retain only the top-K most relevant connections per query token to reduce background contamination while preserving subject identity. We further propose Next-Step Variance Ascent (NSVA), which dynamically adjusts K across denoising steps by maximizing between-class attention variance, and foreground-background separation to restrict cross-image attention exclusively to subject-relevant tokens. Experiments on the ConsiStory+ benchmark demonstrate that SelectDiff outperforms both fine-tuning-based and training-free methods in identity preservation and prompt alignment, while achieving superior inference efficiency. As a natural byproduct, our approach handles multi-subject scenarios without any specialized processing.

1 INTRODUCTION

Text-to-image diffusion models (Rombach et al., 2022; Wang et al., 2025c) generate high-quality images, but preserving subject appearance across scenes remains challenging (Avrahami et al., 2024; Yuan et al., 2025; Wang et al., 2025a). Such consistency is essential for personalized generation (Feng et al., 2025; Nguyen et al., 2025), storytelling (Zhou et al., 2024; Liu et al., 2025b; Yang et al., 2025), and multi-view product visualization (Yang et al., 2024; Chen et al., 2024).

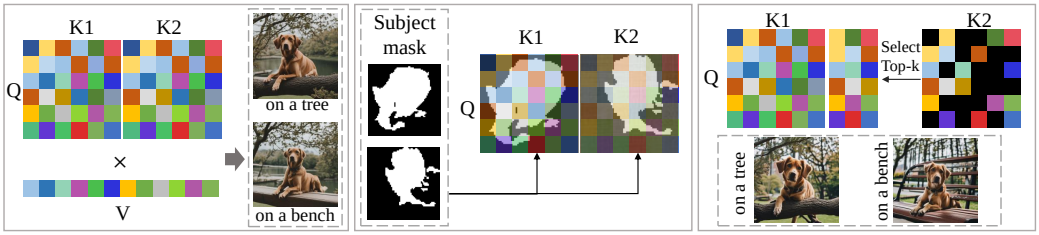
Subject-specific fine-tuning (Zhou et al., 2026; Yang et al., 2025) encodes identity by optimizing model parameters or embeddings on reference images, but requires costly optimization for each new subject. Training-free alternatives instead operate at inference time.

Inspired by video generation (Wu et al., 2023), training-free methods (Liu et al., 2025b; Tewel et al., 2024) share self-attention features across images to promote subject consistency. However, unconstrained sharing can also propagate irrelevant background features (Fig. 1(a)). ConsiStory (Tewel et al., 2024) suppresses this leakage using subject masks extracted from cross-attention maps (Fig. 1(b)), requiring additional feature capture and mask computation at every denoising step.

We propose **SelectDiff** (**Select**ive attention-guided consistent **Diff**usion), a training-free framework that exploits the higher similarity of semantically corresponding tokens. Retaining only the top-K connections per query promotes subject consistency, reduces background contamination, and lowers cross-image aggregation cost (Fig. 1(c)). Selective sharing also supports multi-subject generation without specialized processing.

Next-Step Variance Ascent (NSVA) adapts a separate K for each attention head using a local finite difference of the between-class variance of selected and unselected scores. Updates take effect at the next denoising step without additional denoiser forward passes. We also observe that selected top-K index sets stabilize across steps. Caching these indices while recomputing attention with current features avoids repeated full cross-image similarity searches and token selection.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Naive cross-image attention sharing **(b) Mask-guided attention sharing** **(c) Similarity-guided selective attention**

Figure 1: Comparison among different methods. **(a)** Naive cross-image attention fusion leads to background leakage, where irrelevant background features from one image contaminate the target generation. **(b)** ConsiStory (Tewel et al., 2024) extracts subject masks from cross-attention maps to block background feature sharing, at the cost of per-iteration mask computation. **(c)** Our method selectively retains only the top-K most relevant connections per query token, filtering cross-image connections before feature aggregation.

Foreground queries tend to have higher mean top-K similarity scores than background queries. We use this distinction to restrict cross-image interaction to subject-relevant queries, while background queries retain within-image attention, avoiding auxiliary mask extraction.

In summary, our contributions are as follows:

- Focusing on the problem of background leakage, we propose a straightforward yet effective training-free method based on similarity-guided selective attention.
- We propose a Next-Step Variance Ascent (NSVA) mechanism that dynamically adjusts the top-K value in denoising iterations by maximizing attention variance, allowing adaptive selection of semantically relevant tokens while maintaining computational efficiency.
- We further leverage the distinct attention value distributions of foreground and background tokens to explicitly separate them, fundamentally eliminating background leakage.

2 Related Work

2.1 Text-to-Image Generation

Diffusion models (Ho et al., 2020; Jiang et al., 2025) underpin high-quality text-to-image generation. Latent Diffusion Models (Rombach et al., 2022) improve efficiency through latent-space generation, and SDXL (Podell et al., 2024) scales the U-Net backbone for higher fidelity and text alignment. Flow matching (Shin et al., 2025; Esser et al., 2024) offers an alternative training paradigm, as used in Stable Diffusion 3 (Esser et al., 2024) and FLUX (Gao et al., 2025).

For controllable generation, ControlNet (Zhang et al., 2023) and T2I-Adapter (Mou et al., 2024) introduce spatial conditioning, while MS-Diffusion (Wang et al., 2025b) supports layout-guided multi-subject personalization. Our focus is training-free identity consistency across generated images.

2.2 Subject-Consistent Text-to-Image Generation

Subject-consistent generation (Liu et al., 2025a; He et al., 2025; Zhou et al., 2025) includes test-time fine-tuning methods such as DreamBooth (Ruiz et al., 2023), which encode identity from reference images. For human-centric stories, IdentityStory (Zhou et al., 2026) combines iterative identity discovery with re-denoising identity injection. Subject-specific optimization, however, limits scalability.

Methods without test-time optimization include IP-Adapter (Ye et al., 2023), which injects visual prompts through decoupled cross-attention, and JeDi (Zeng et al., 2024), which uses joint-image diffusion. StoryDiffusion (Zhou et al., 2024) shares self-attention across images, while 1Prompt1Story (Liu et al., 2025b) combines concatenated prompts with singular-value-based semantic modulation. Con-

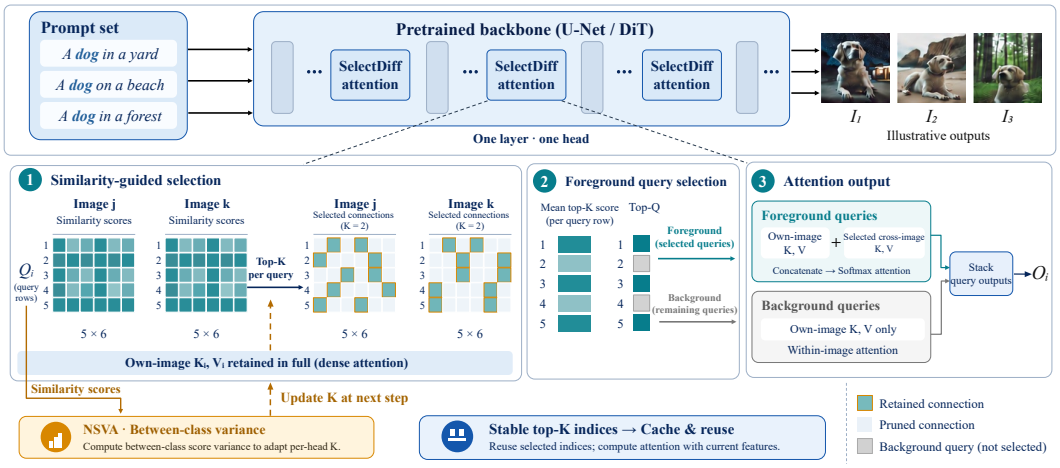


Figure 2: Overview of SelectDiff. At each denoising step, SGSA retains only the top-K most relevant cross-image token connections per query. The mean top-K similarity score is then used to separate foreground from background, restricting cross-image attention to the top-Q foreground query tokens. NSVA adaptively updates K across steps by maximizing between-class variance.

siStory (Tewel et al., 2024) restricts feature sharing with subject masks computed at each step. SelectDiff instead selects relevant cross-image token connections without auxiliary mask extraction.

3 METHODS

3.1 OVERVIEW

Given N prompts $\{p_i\}_{i=1}^N$, SelectDiff jointly generates subject-consistent images $\{\mathcal{I}_i\}_{i=1}^N$ without training or test-time optimization. At each denoising step, image $\mathcal{I}_i$ supplies queries, keys, and values $\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i$. Similarity-guided selective attention (Section 3.2) retains the top-K connections from each other image per query. NSVA (Section 3.3) adapts K for the next step without additional denoiser forward passes.

Foreground-background separation (Section 3.4) selects the top-Q queries by mean top-K similarity and permits cross-image attention only for these queries; remaining queries use within-image attention. Fig. 2 summarizes the pipeline.

Selection is performed independently in each attention head at each layer, allowing different heads to use different cardinalities. Within a head, each query selects its own reference tokens, preserving query-specific correspondences across images. Own-image keys and values remain available in full, providing local context alongside the selected cross-image features. As denoising proceeds, we exploit the observed stability of selected index sets to cache the connectivity pattern. Attention scores and aggregated values are still computed from current features. This separates reusable token correspondences from evolving feature representations; the caching rationale depends on index stability, rather than an unchanged value of K alone.

Notably, our method naturally maintains consistency across multiple subjects within a single image, even for subjects of the same category. When multiple subjects of the same category appear across images, tokens from each subject primarily attend to their semantically corresponding tokens in other images, which produce the highest similarity scores. Tokens from different subjects exhibit lower cross-subject similarity due to spatial and contextual differences and are therefore naturally filtered out by the top-K selection.

3.2 SIMILARITY-GUIDED SELECTIVE ATTENTION

We sparsify cross-image feature aggregation by selecting the top-K keys and values from each other image for each query. Selection still requires query–key similarity computation over the candidate

tokens; sparsity reduces the subsequent aggregation cost, rather than eliminating the cost of finding the selected connections. We describe one attention head at one layer and denoising step, suppressing the layer, head, and timestep indices. Here K is the cardinality maintained by that head. For each target image $\mathcal{I}_i$, let $\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i \in \mathbb{R}^{L\times d}$, where $L = HW$ is the number of spatial tokens per image, d is the head dimension, and N is the number of images. For a fixed target image i, write $\mathbf{q}_n = \mathbf{Q}_i[n,:] \in \mathbb{R}^{1\times d}$. We compute its similarity to the keys of each other image $\mathcal{I}_j$ ($j \neq i$):

$$\mathbf{s}_{n,j} = \mathbf{q}_n \mathbf{K}_j^\top, \tag{1}$$

where $\mathbf{s}_{n,j} \in \mathbb{R}^{1\times L}$. Then, we select the top-K indices for each other image:

$$\mathcal{T}_{n,j} = \underset{\mathcal{T}\subseteq\{1,\dots,L\},\ |\mathcal{T}|=K}{\arg\max} \sum_{r\in\mathcal{T}} s_{n,j,r}. \tag{2}$$

The target-image index i is implicit in $\mathbf{s}_{n,j}$ and $\mathcal{T}_{n,j}$. Each query has its own selection sets, even though the cardinality K is shared within the head. For query n, we concatenate its selected keys and values from the other images:

$$\tilde{\mathbf{K}}_{i,n} = \bigoplus_{j\neq i} \mathbf{K}_j[\mathcal{T}_{n,j}], \quad \tilde{\mathbf{V}}_{i,n} = \bigoplus_{j\neq i} \mathbf{V}_j[\mathcal{T}_{n,j}], \tag{3}$$

where $\bigoplus$ denotes concatenation along the token dimension and $\mathbf{K}_j[\mathcal{T}_{n,j}]$ denotes the rows of $\mathbf{K}_j$ indexed by $\mathcal{T}_{n,j}$. Thus, $\tilde{\mathbf{K}}_{i,n}, \tilde{\mathbf{V}}_{i,n} \in \mathbb{R}^{(N-1)K\times d}$ are query-dependent matrices. The output for this query is

$$\mathbf{o}_{i,n}^{\text{sel}} = \text{softmax}\left(\frac{\mathbf{q}_n[\mathbf{K}_i;\ \tilde{\mathbf{K}}_{i,n}]^\top}{\sqrt{d}}\right)[\mathbf{V}_i;\ \tilde{\mathbf{V}}_{i,n}], \tag{4}$$

where $[\cdot\,;\cdot]$ denotes concatenation along the token dimension and softmax is taken over the $L + (N-1)K$ keys available to this query. Each $\mathbf{o}_{i,n}^{\text{sel}} \in \mathbb{R}^{1\times d}$ is a row; stacking these rows yields $\mathbf{O}_i^{\text{sel}} = [\mathbf{o}_{i,1}^{\text{sel}}; \dots; \mathbf{o}_{i,L}^{\text{sel}}] \in \mathbb{R}^{L\times d}$. This row-wise definition permits batched implementation with query-dependent gathered keys and values, without assuming a shared selected matrix for all queries. Note that the target image's own keys $\mathbf{K}_i$ and values $\mathbf{V}_i$ are included in full without top-K filtering, preserving the integrity of within-image self-attention while restricting cross-image interaction to the most semantically relevant tokens.

The key intuition behind Eq. equation 2 is that tokens belonging to the same semantic category naturally exhibit higher mutual similarity. Foreground subject tokens in $\mathcal{I}_i$ share similar visual semantics with corresponding subject tokens in other images, producing a larger dot-product similarity scores. Background tokens, by contrast, are semantically unrelated across images and yield consistently lower scores. Therefore, top-K selection naturally biases toward subject-to-subject connections, ensuring that aggregated features predominantly encode subject identity rather than irrelevant background content.

3.3 Next-Step Variance Ascent

While the top-K selective attention effectively suppresses background leakage, the optimal value of K varies across different images and subjects. Manually tuning K for each input is impractical. We therefore propose *Next-Step Variance Ascent* (NSVA), a lightweight mechanism that automatically adapts K throughout the denoising process without additional forward passes.

Each attention layer ℓ maintains a separate selection cardinality $K_t^{(\ell,h)}$ for each head h; cardinalities are not shared across heads or layers. Within a head, each query still selects its own reference-token indices using that head's cardinality. Below, we suppress (ℓ, h) for readability and first define the score statistic for an individual query–reference pair.

We use between-class variance as a surrogate for separation between high- and low-similarity connections, rather than assuming that its maximizer is also optimal for identity preservation. For one query and one reference image at timestep t, let $s_{t,(1)} \geq \cdots \geq s_{t,(L)}$ be the sorted raw query–key scores from Eq. equation 1; query, image, layer, and head indices are suppressed here. For an integer $1 \leq K \leq L-1$, define

$$\mu_t^+(K) = \frac{1}{K}\sum_{r=1}^{K} s_{t,(r)}, \qquad \mu_t^-(K) = \frac{1}{L-K}\sum_{r=K+1}^{L} s_{t,(r)}. \tag{5}$$

Following the between-class variance criterion used in Otsu thresholding (Otsu, 1979), we define

$$\mathrm{Var}_t(K) = \frac{K}{L} \cdot \frac{L-K}{L} \cdot \left(\mu_t^+(K) - \mu_t^-(K)\right)^2 . \tag{6}$$

This objective measures score separation, not semantic labels, and does not guarantee that selected and unselected tokens correspond to different subjects or to foreground and background. Once the scores are sorted, prefix sums allow all candidate variances to be evaluated in $O(L)$ additional work per score vector. NSVA instead uses a local finite difference and carries the update to the next denoising step, avoiding a full search over candidate cardinalities at each step. At an interior value $1 \le K_t \le L-2$, this difference is

$$g_t = \mathrm{Var}_t(K_t + 1) - \mathrm{Var}_t(K_t), \tag{7}$$

and the unquantized update proposal for the next denoising step is

$$\widehat{K}_{t-1} = K_t + \eta g_t, \tag{8}$$

where $\eta > 0$ is the update step size. The proposal must be converted to an integer cardinality in $[1, L-1]$ before token selection, and the boundary case $K_t = L-1$ requires a one-sided difference that remains in this domain. The two variance evaluations can reuse the total score sum, the top-K_t score sum, and the (K_t+1)-th score. Obtaining these statistics still has a selection and reduction cost, but requires no additional denoiser forward pass. Because the score distribution changes with t and the objective need not be unimodal, this local update does not guarantee monotonic ascent across steps or convergence to a global optimum. The trajectories in the appendix provide empirical evidence of stabilization for the illustrated inputs only.

Caching temporally stable top-K indices. We empirically observe that the selected top-K index sets become stable across denoising steps. This observation concerns the identities of the selected tokens, rather than only the cardinality $K_t^{(\ell,h)}$. We exploit this temporal stability by caching the query-specific index sets within each attention head and reusing them in subsequent steps. Attention on the cached connections is computed using the current query, key, and value features, so feature values and attention weights continue to evolve while repeated full candidate searches and top-K selection are avoided. The appendix separates the cost of constructing the index sets from the cost of reusing them.

3.4 Foreground-Background Separation

Although top-K selective attention filters irrelevant tokens from other images for each individual query, background query tokens in the target image still participate in cross-image feature interaction. Since background regions are semantically unrelated to the subject, allowing them to attend to features from other images—even selectively—introduces unnecessary noise. We therefore propose to explicitly identify and exclude background query tokens from cross-image attention entirely.

For each query token $\mathbf{q}_n$ in the target image $\mathcal{I}_i$, we define its foreground score f_n as the mean attention similarity to its top-K most similar tokens selected from other images, using the selected-score mean defined in Eq. equation 5.

A higher f_n indicates that the query token is semantically close to the subject tokens in other images, and is therefore more likely to belong to the foreground. Given foreground scores $\{f_n\}_{n=1}^{L}$, we identify foreground query tokens by selecting the top-Q queries with the highest scores:

$$\mathcal{F}_i = \underset{|\mathcal{F}|=Q}{\arg\max} \{ f_n \mid n = 1, \ldots, L \} . \tag{9}$$

Cross-image feature sharing is then restricted exclusively to foreground query tokens:

$$\mathbf{o}_{i,n} = \begin{cases} \mathbf{o}_{i,n}^{\mathrm{sel}} & \text{if } n \in \mathcal{F}_i, \\ \mathrm{SelfAttn}(\mathbf{q}_n,\ \mathbf{K}_i,\ \mathbf{V}_i) & \text{otherwise}, \end{cases} \tag{10}$$

where $\mathbf{o}_{i,n}^{\mathrm{sel}}$ is defined in Eq. equation 4 using the selected keys and values for query n. The final per-head output is $\mathbf{O}_i = [\mathbf{o}_{i,1}; \ldots; \mathbf{o}_{i,L}]$. Background tokens fall back to standard within-image self-attention, ensuring they are never contaminated by features from other images.

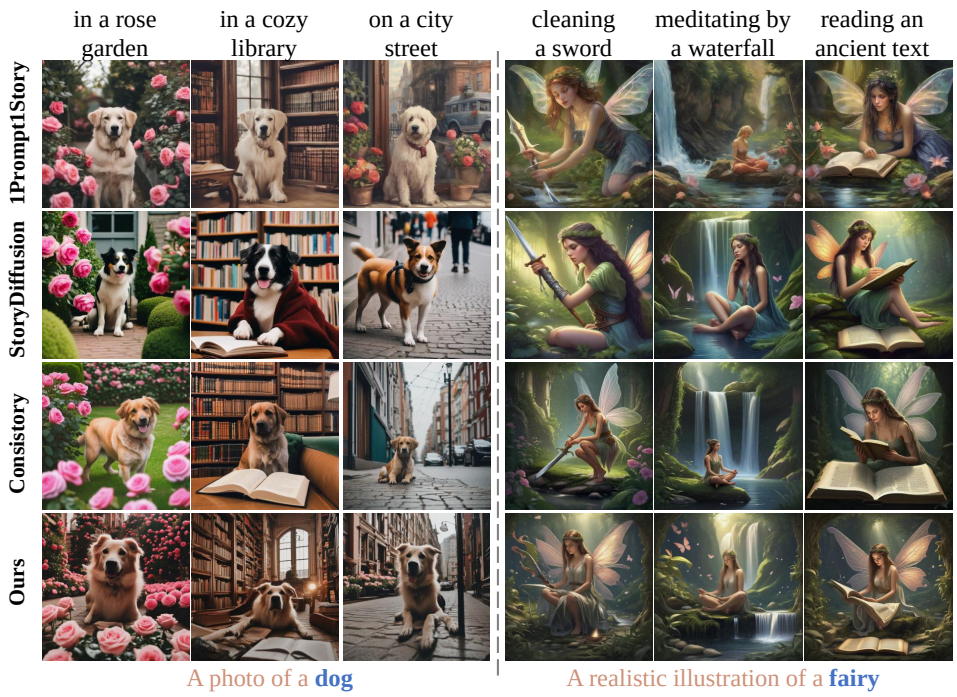


Figure 3: Qualitative comparison of single-subject consistent generation across different scenes. We show two example subjects: a dog (left) and a fairy (right). For each subject, images are generated under varying scene prompts. 1Prompt1Story and StoryDiffusion fail to maintain a consistent subject appearance in the dog example. Our method SelectDiff achieves the best identity preservation across all scenes for both subjects.

Our approach differs fundamentally from ConsiStory (Tewel et al., 2024), which extracts subject masks from cross-attention maps at every denoising iteration, requiring additional computation per step. In contrast, our foreground scores $\{f_n\}$ are derived directly from the already-computed top-K similarity scores in Eq. equation 2, introducing zero additional computation. Furthermore, our method operates entirely in the self-attention feature space without relying on text-conditioned cross-attention maps, making it applicable to any diffusion backbone regardless of architecture.

4 Experiments

4.1 Experimental Setup

We evaluate SelectDiff against representative methods from both fine-tuning-based and training-free categories, including DreamBooth (Ruiz et al., 2023), PhotoMaker (Li et al., 2024), IP-Adapter (Ye et al., 2023), naive Cross-image Self-Attention (CSA), ConsiStory (Tewel et al., 2024), StoryDiffusion (Zhou et al., 2024), and 1p1s (Liu et al., 2025b). Following the benchmark proposed in 1p1s, we adopt ConsiStory+, a challenging evaluation protocol that generates up to 1,500 images across 200

Table 1: Quantitative comparison with different methods. ↑ indicates higher is better, ↓ indicates lower is better. Best results are in **bold**, second best are underlined.

Method	Backbone	Training-Free	CLIP-T ↑	CLIP-I ↑	DreamSim ↓	Memory (GB) ↓	Time (s) ↓
PhotoMaker (Li et al., 2024)	SDXL	✗	86.5	84.7	39.9	23.7	18.0
IP-Adapter (Ye et al., 2023)	SDXL	✗	84.6	94.3	14.6	19.4	13.5
Naive CSA	SDXL	✓	72.5	74.2	33.8	42.2	38.0
ConsiStory (Tewel et al., 2024)	SDXL	✓	87.7	87.4	31.9	34.5	34.6
StoryDiffusion (Zhou et al., 2024)	SDXL	✓	88.7	87.5	32.1	45.6	25.7
1Prompt1Story (Liu et al., 2025b)	SDXL	✓	89.4	91.2	19.9	18.7	23.2
SelectDiff (Ours)	SDXL	✓	90.2	92.7	18.5	22.6	19.5
SelectDiff+ (Ours)	DiT	✓	**92.1**	**94.7**	**13.2**	30.2	24.6



Figure 4: Qualitative comparison of multi-subject consistent generation with increasing difficulty: (a) semantically distinct instances, (b) similar instances (origami-style hedgehog and dog), and (c) watercolor style. Our method maintains clear identity separation in (a) and (b), where ConsiStory already fails in (b). All methods struggle in (c).

prompt sets, covering a diverse range of subjects and scenes. All experiments are conducted on an NVIDIA A100 GPU. To assess computational efficiency, we report inference time in seconds (s) and memory consumption in gigabytes (GB), denoted as **Time** and **Memory**, respectively.

Following (Liu et al., 2025b), we compute the average CLIP score (Hessel et al., 2021) between each generated image and its corresponding prompt, denoted as **CLIP-T**. For identity consistency, we employ **CLIP-I** (Hessel et al., 2021), which measures cosine similarity between CLIP image embeddings, and **DreamSim** (Fu et al., 2023), which closely aligns with human perceptual judgment.

4.2 Results

Quantitative Comparison. Table 1 presents the quantitative comparison of SelectDiff against both training-based and training-free methods. Among all methods, SelectDiff achieves the highest CLIP-T score of 90.2, surpassing all baselines including training-based approaches, indicating superior prompt alignment. In terms of identity consistency, our method ranks second overall with a CLIP-I of 92.7 and DreamSim of 18.5, behind only IP-Adapter (Ye et al., 2023). However, the high identity similarity of IP-Adapter largely stems from its tendency to generate images with subjects in similar poses and layouts, reflecting identity overfitting rather than genuine appearance consistency. This is further corroborated by its relatively low CLIP-T score of 84.6, suggesting poor prompt alignment.

Among training-free methods, SelectDiff achieves the best overall balance across all metrics. Compared to ConsiStory (Tewel et al., 2024) and StoryDiffusion (Zhou et al., 2024), our method improves identity consistency by a significant margin while also achieving better prompt alignment. Notably, compared to the naive cross-image self-attention baseline (Naive CSA), our method substantially improves all metrics, directly validating the effectiveness of our similarity-guided selective attention in suppressing background leakage. In terms of efficiency, SelectDiff achieves the fastest inference time and low memory consumption among training-free methods, showing an empirical runtime benefit for the evaluated implementation. This measurement includes selection and attention computation; sparsity of the retained connections alone does not imply the same reduction in total arithmetic cost.

Beyond the SDXL-based comparison above, we explore the applicability of SelectDiff to DiT-based generation using FLUX.2 [klein] 4B Base (Black Forest Labs, 2026), an undistilled 4-billion-parameter rectified-flow transformer, as an additional backbone. The extension applies selective cross-image connections to image tokens while retaining the original text conditioning, without updating model parameters. The SelectDiff (FLUX.2) row in Table 1 is reserved for this evaluation under the ConsiStory+ protocol, using the same quality metrics and reporting time and memory in seconds and GB, respectively.

Qualitative Comparison of Single-Subject Consistent. Fig. 3 presents qualitative comparisons between SelectDiff and training-free baselines including ConsiStory (Tewel et al., 2024), StoryDiffusion (Zhou et al., 2024), and 1Prompt1Story (Liu et al., 2025b) across two example subjects: a dog and a fairy. As shown, 1Prompt1Story and StoryDiffusion fail to maintain consistent subject appearance in the dog example, producing images where the subject's identity noticeably varies across different scenes. In contrast, SelectDiff consistently preserves the finest details of subject appearance across all generated images in both examples, demonstrating the effectiveness of our similarity-guided selective attention in retaining identity-relevant features while suppressing irrelevant background contamination.

Qualitative Comparison of Multi-Subject Consistent. A notable property of SelectDiff is its natural handling of multi-instance scenarios without any specialized processing, as shown in Fig. 4. 1Prompt1Story (Liu et al., 2025b) and StoryDiffusion (Zhou et al., 2024) fail entirely at preserving multiple distinct identities. ConsiStory (Tewel et al., 2024) partially handles semantically distinct instances, as the exponential form of softmax acts as a gating mechanism. However, when instances are semantically similar, this gating is insufficient. Given an origami-style hedgehog and a dog, ConsiStory generates an ambiguous figure blending both subjects. Our method maintains clear identity separation in both cases. In the third group, all methods fail under watercolor style, where weakened semantic distinctiveness prevents reliable instance-level feature matching, suggesting style-robust multi-instance generation remains an open problem.

User Study. To further validate the perceptual quality of our results, we conduct a user study via an online questionnaire distributed to 100 participants. For each prompt set, participants are presented with generated image sequences from different methods in randomized order and asked to rate them on identity consistency, prompt alignment, and image diversity. Results in Fig. 5 demonstrate that SelectDiff is strongly preferred by human evaluators on both criteria.

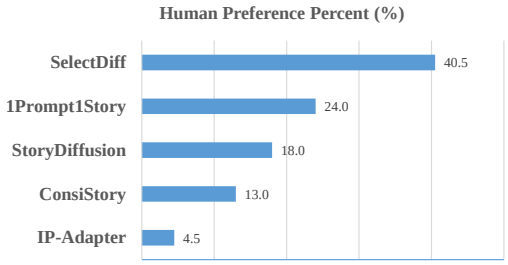


Figure 5: User study with 100 people to vote for the best consistent generation method according to human preference.

4.3 Analysis

Component Ablation. We conduct ablation studies to validate the contribution of each component in SelectDiff. All ablations are evaluated on the ConsiStory+ benchmark. Table 2 reports the ablation results for each component of SelectDiff, i.e., similarity-guided selective attention (SGSA), Next-Step Variance Ascent (NSVA), and Foreground-Background Separation (FBS). When removing SGSA, we use naive cross-image attention fusion to achieve subject consistency.

SGSA provides the largest improvement over naive attention fusion while substantially reducing memory and runtime. NSVA further improves identity consistency with negligible time overhead relative to the evaluated fixed-K setting, without establishing an optimal semantic boundary. FBS improves quality and reduces runtime by restricting cross-image attention to foreground queries. The full model achieves the best quality scores, supporting the components' complementary roles.

Effect of Top-K in SGSA and Top-Q in FBS. We investigate the two key hyperparameters in SelectDiff: the number of selected reference tokens K in SGSA and foreground query tokens Q in FBS. As shown in Fig. 6, increasing K improves subject identity preservation but introduces background leakage, while decreasing K suppresses contamination at the cost of identity consistency. Similarly, a larger Q benefits subject consistency but risks background contamination—at Q-rate $=$

Table 2: The performance of different ablation settings.

SGSA	NSVA	FBS	CLIP-T ↑	CLIP-I ↑	DreamSim ↓	Memory ↓ (GB)	Time ↓ (s)
×	×	×	72.5	74.2	33.8	42.2	38.0
√	×	×	83.1	84.3	26.8	22.6	22.5
√	√	×	86.4	87.2	22.9	22.6	22.7
×	×	√	75.3	77.6	30.9	42.2	34.6
√	×	√	85.2	86.8	24.3	22.6	19.4
√	√	√	**90.2**	**92.7**	**18.5**	22.6	19.5



Figure 6: Visual analysis of Top-K-rate (rows) and Top-Q-rate (columns) using the prompt **"a photo of a dog"** with scenes **"in a ship"** and **"in a garden"**.

0.8, K-rate $= 0.5$, severe leakage is observed—whereas a smaller Q may miss subject tokens near object boundaries. We initialize K at approximately 1% of the tokens in each reference image, with the cardinality subsequently adapting per input. This visual comparison illustrates the trade-off controlled by K and Q; it does not establish that NSVA is preferable to every fixed setting.

5 Conclusion

In this paper, we present SelectDiff, a training-free framework for subject-consistent image generation. Unconstrained cross-image feature aggregation can introduce irrelevant background information. To mitigate this issue, we propose similarity-guided selective attention that retains only the top-K most relevant token connections. We further introduce NSVA to adaptively adjust K via between-class variance maximization, and foreground-background separation to restrict cross-image attention to subject-relevant tokens. Experiments on the ConsiStory+ benchmark demonstrate that SelectDiff outperforms both fine-tuning-based and training-free baselines in identity consistency and prompt alignment, while achieving superior inference efficiency and naturally handling multi-instance scenarios without specialized processing.

AI USE STATEMENT

In this work, we used generative AI tools for grammar correction of the manuscript text and typographical error checking only. We have not used generative AI tools for all other mandatory disclosure scenarios including generating core research ideas, deriving theoretical conclusions, drafting core experimental code, producing original experimental results, designing study protocols and compiling reference lists, and all remaining mandatory disclosure categories related to core research content generation are not applicable to this work. Additionally, we used generative AI tools for no extra tasks that fall under the recommended disclosure scope. We have reviewed all AI-assisted work. We manually cross-checked every AI-adjusted text segment by all co-authors, to confirm that the generative AI tools only fixed grammatical mistakes and corrected typos, without altering any of our original academic claims, technical descriptions, experimental conclusions and logical flow of the manuscript, no unintended semantic deviation or information loss was introduced during this process. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

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. In *Proceedings of ACM SIGGRAPH conference*, pp. 1–12, 2024.

Black Forest Labs. FLUX.2 [klein] 4B Base. `https://huggingface.co/black-forest-labs/FLUX.2-klein-base-4B`, 2026. Official model card.

Binghui Chen, Chongyang Zhong, Wangmeng Xiang, Yifeng Geng, and Xuansong Xie. Virtualmodel: Generating object-id-retentive human-object interaction image by diffusion model for e-commerce marketing. *arXiv preprint arXiv:2405.09985*, 2024.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the International Conference on Machine Learning*, 2024.

Haoran Feng, Zehuan Huang, Lin Li, Hairong Lv, and Lu Sheng. Personalize anything for free with diffusion transformer. *arXiv preprint arXiv:2503.12590*, 2025.

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. In *Proceedings of the Advances in Neural Information Processing Systems*, 2023.

Daiheng Gao, Shilin Lu, Wenbo Zhou, Jiaming Chu, Jie Zhang, Mengxi Jia, Bang Zhang, Zhaoxin Fan, and Weiming Zhang. Eraseanything: Enabling concept erasure in rectified flow transformers. In *Proceedings of the International Conference on Machine Learning*, 2025.

Huiguo He, Huan Yang, Zixi Tuo, Yuan Zhou, Qiuyue Wang, Yuhang Zhang, Zeyu Liu, Wenhao Huang, Hongyang Chao, and Jian Yin. Dreamstory: Open-domain story visualization by llm-guided multi-subject consistent diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the conference on Empirical Methods in Natural Language Processing*, pp. 7514–7528, 2021.

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

Wan Jiang, He Wang, Xin Zhang, Dan Guo, Zhaoxin Fan, Yunfeng Diao, and Richang Hong. Moderating the generalization of score-based generative model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 360–369, 2025.

Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8640–8650, 2024.

Lijie Liu, Tianxiang Ma, Bingchuan Li, Zhuowei Chen, Jiawei Liu, Gen Li, Siyu Zhou, Qian He, and Xinglong Wu. Phantom: Subject-consistent video generation via cross-modal alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14951–14961, 2025a.

Tao Liu, Kai Wang, Senmao Li, Joost van de Weijer, Fhad Khan, Shiqi Yang, Yaxing Wang, Jian Yang, and Mingming Cheng. One-prompt-one-story: Free-lunch consistent text-to-image generation using a single prompt. In *Proceedings of the International Conference on Learning Representations*, 2025b.

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying 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*, volume 38, pp. 4296–4304, 2024.

Thao Nguyen, Krishna Kumar Singh, Jing Shi, Trung Bui, Yong Jae Lee, and Yuheng Li. Yo'chameleon: Personalized vision and language generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14438–14448, 2025.

Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979. doi: 10.1109/TSMC.1979.4310076.

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. In *Proceedings of the International Conference on Learning Representations*, pp. 1862–1874, 2024.

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

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.

Inkyu Shin, Chenglin Yang, and Liang-Chieh Chen. Deeply supervised flow-based generative models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16535–16544, 2025.

Yoad Tewel, Omri Kaduri, Rinon Gal, Yoni Kasten, Lior Wolf, Gal Chechik, and Yuval Atzmon. Training-free consistent text-to-image generation. *ACM Trans. Graph.*, 43(4):52:1–52:18, 2024.

Qinghe Wang, Baolu Li, Xiaomin Li, Bing Cao, Liqian Ma, Huchuan Lu, and Xu Jia. Characterfactory: Sampling consistent characters with gans for diffusion models. *IEEE Transactions on Image Processing*, 2025a.

Xierui Wang, Siming Fu, Qihan Huang, Wanggui He, and Hao Jiang. MS-diffusion: Multi-subject zero-shot image personalization with layout guidance. In *Proceedings of the International Conference on Learning Representations*, 2025b.

Zhendong Wang, Jianmin Bao, Shuyang Gu, Dong Chen, Wengang Zhou, and Houqiang Li. Design-diffusion: High-quality text-to-design image generation with diffusion models. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20906–20915, 2025c.

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.

Jiayu Yang, Ziang Cheng, Yunfei Duan, Pan Ji, and Hongdong Li. Consistnet: Enforcing 3d consistency for multi-view images diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7079–7088, 2024.

Shuai Yang, Yuying Ge, Yang Li, Yukang Chen, Yixiao Ge, Ying Shan, and Ying-Cong Chen. Seed-story: Multimodal long story generation with large language model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1850–1860, 2025.

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

Shenghai Yuan, Jinfa Huang, Xianyi He, Yunyang Ge, Yujun Shi, Liuhan Chen, Jiebo Luo, and Li Yuan. Identity-preserving text-to-video generation by frequency decomposition. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12978–12988, 2025.

Yu Zeng, Vishal M Patel, Haochen Wang, Xun Huang, Ting-Chun Wang, Ming-Yu Liu, and Yogesh Balaji. Jedi: Joint-image diffusion models for finetuning-free personalized text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6786–6795, 2024.

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

Donghao Zhou, Jingyu Lin, Guibao Shen, Quande Liu, Jialin Gao, Lihao Liu, Lan Du, Cunjian Chen, Chi-Wing Fu, Xiaowei Hu, and Pheng-Ann Heng. Identitystory: Taming your identity-preserving generator for human-centric story generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

Tianchen Zhou, Zhongjie Duan, Cen Chen, Wenmeng Zhou, Yanhao Wang, and Yaliang Li. Agentstory: A multi-agent system for story visualization with multi-subject consistent text-to-image generation. In *Proceedings of the International Conference on Multimedia Retrieval*, pp. 1894–1902, 2025.

Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation. *Proceedings of Advances in Neural Information Processing Systems*, 37:110315–110340, 2024.