ADASHARE: ADAPTIVE CROSS-FRAME SHARING FOR TRAINING-FREE CONSISTENT IMAGE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Consistent visual story generation requires maintaining subject identity and visual style across images under different scene prompts. Existing training-free methods improve cross-frame consistency by sharing reference features across frames. However, unreliable feature sharing can lead to incomplete subject generation and degraded image quality. To address these issues, we propose a training-free framework for adaptive cross-frame feature sharing in consistent text-to-image generation, termed AdaShare. To improve subject completeness and consistency, AdaShare extracts reliable subject masks from cross-attention maps and activates subject feature sharing after subject regions stabilize. We further introduce *Adaptive-Capacity Optimal Transport Attention* (*ACOT Attention*), which formulates *Extended Attention* as an optimal transport problem. By imposing adaptive capacity constraints derived from the areas of the subject masks, *ACOT Attention* prevents excessive attention concentration on reference tokens caused by token imbalance, enabling more effective utilization of cross-frame features. Furthermore, to maintain style consistency, we introduce *Covariance Value Alignment*, which applies whitening and coloring transformation to the low-frequency components of value representations while leaving high-frequency structures unchanged. Extensive experiments on ConsiStory+ demonstrate that AdaShare achieves superior performance over training-free baselines in core consistency metrics and human evaluations, while maintaining prompt fidelity and image quality. The source code will be released upon acceptance.

1 INTRODUCTION

Consistent text-to-image generation (Tewel et al., 2024) aims to preserve subject identity and appearance consistency across scenes specified by different text prompts. The subject should remain visually consistent while appropriately adapting to the scene and context described by each prompt. Furthermore, the generated image sequence should exhibit consistent visual style. Existing methods (Ruiz et al., 2023; Wu et al., 2025; Kumari et al., 2023) rely on model training or additional fine-tuning, which introduces considerable training overhead and limits their applicability.

Recently, training-free methods (Zhou et al., 2024; Gao et al., 2026) have attracted increasing attention. These approaches directly leverage feature representations from pretrained generative models without modifying model parameters. For example, ConsiStory (Tewel et al., 2024) propagates subject features across generated images through subject-driven shared attention and feature injection. 1Prompt1Story (Liu et al., 2025) exploits the consistency of linguistic context and identity-preserving cross-attention to maintain subject identity. Infinite-Story (Park et al., 2026) further incorporates both identity and style consistency into a training-free framework. Despite these advances, cross-image feature sharing in diffusion-based methods still poses significant challenges.

First, subject masks used to guide feature sharing are typically derived from cross-attention maps (Gao et al., 2026; Hertz et al., 2023). However, attention responses across different layers and denoising steps often suffer from spatial diffusion. Directly aggregating these unreliable maps introduces noise into the subject masks, resulting in inaccurate cross-image feature interactions. Second, conventional *Extended Attention* (Tewel et al., 2024) applies row-wise softmax, which independently normalizes the attention distribution for each query but lacks a global constraint over the reference tokens. Consequently, multiple queries may concentrate on reference tokens with strong responses, which can further cause incomplete subject and degraded image quality. Third, to achieve style consistency, directly applying the whitening and coloring transformation (WCT) (Li

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.

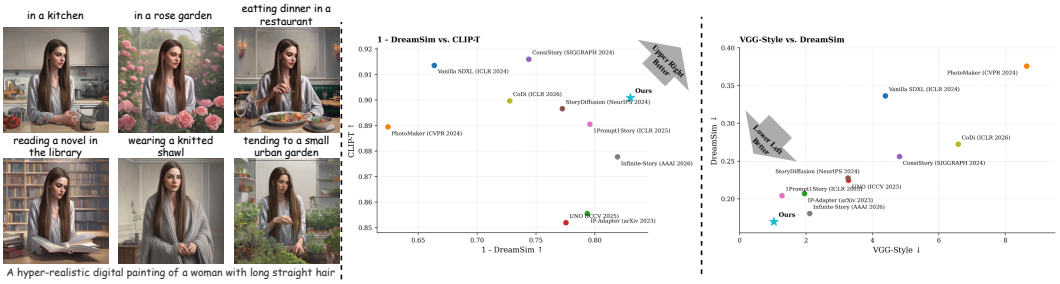


Figure 1: **(Left)** Generated results from our AdaShare. **(Middle)** Comparison of subject consistency measured by 1-DreamSim and text fidelity measured by CLIP-T. **(Right)** Comparison of style consistency measured by VGG-Style and subject consistency measured by DreamSim.

et al., 2017; Chiu, 2019) to style-encoded value features can perturb their high-frequency components (Qin et al., 2025), thereby compromising overall image quality.

To address these issues, we propose a training-free framework for adaptive cross-frame feature sharing for consistent text-to-image generation, termed AdaShare. First, to obtain reliable subject masks, we introduce an *Adaptive Mask Extraction* module. This module filters out cross-attention maps exhibiting severe spatial dispersion and aggregates the remaining responses to yield robust subject masks. Furthermore, subject feature sharing is activated when subject masks extracted from consecutive denoising steps become stable, avoiding premature feature interaction caused by unreliable masks. Second, we propose *Adaptive-Capacity Optimal Transport Attention* (*ACOT Attention*) to improve cross-frame subject feature sharing. *ACOT Attention* introduces an adaptive capacity constraint for reference tokens, preventing excessive attention mass from concentrating on highly activated reference features. Third, we further introduce *Covariance Value Alignment* (CVA) to maintain visual style consistency across multiple images. CVA employs a *Frequency-Selective Whitening and Coloring Transformation* (FS-WCT), which aligns reference statistics only for the low-frequency value features while preserving the original high-frequency representations. As shown in Fig. 1, our method achieves a better trade-off among subject consistency, text fidelity, and style consistency. In summary, the main contributions of this paper are:

- We propose AdaShare, a training-free framework for consistent text-to-image generation that achieves subject and style consistency through adaptive cross-frame feature sharing.
- For subject consistency, we introduce *Adaptive Mask Extraction* and *ACOT Attention* to enable reliable and adaptive cross-frame feature sharing. The former filters cross-attention responses with spatial diffusion and activates cross-frame feature sharing after subject localization becomes stable. The latter imposes adaptive capacity constraints on reference tokens to alleviate excessive attention concentration and facilitate cross-frame feature interaction.
- For style consistency, we propose CVA with FS-WCT, which aligns reference statistics on low-frequency value features while preserving high-frequency representations.
- Extensive experiments demonstrate that AdaShare provides a favorable balance among subject consistency, style consistency, text alignment, and overall generation quality.

2 RELATED WORK

T2I Generation. Text-to-image (T2I) generation aims to synthesize high-fidelity images that are semantically aligned with textual descriptions. Diffusion models generate high-quality images through a denoising process guided by textual conditioning (Ho et al., 2020; Nichol et al., 2022; Ramesh et al., 2022; Saharia et al., 2022). Latent Diffusion Models (LDMs) (Rombach et al., 2022) perform denoising in a compressed latent space, reducing computational cost and enabling large-scale text-to-image models such as Stable Diffusion and SDXL (Podell et al., 2024). However, conventional T2I models typically generate each image independently, making it challenging to preserve subject identity and visual style across different prompts.

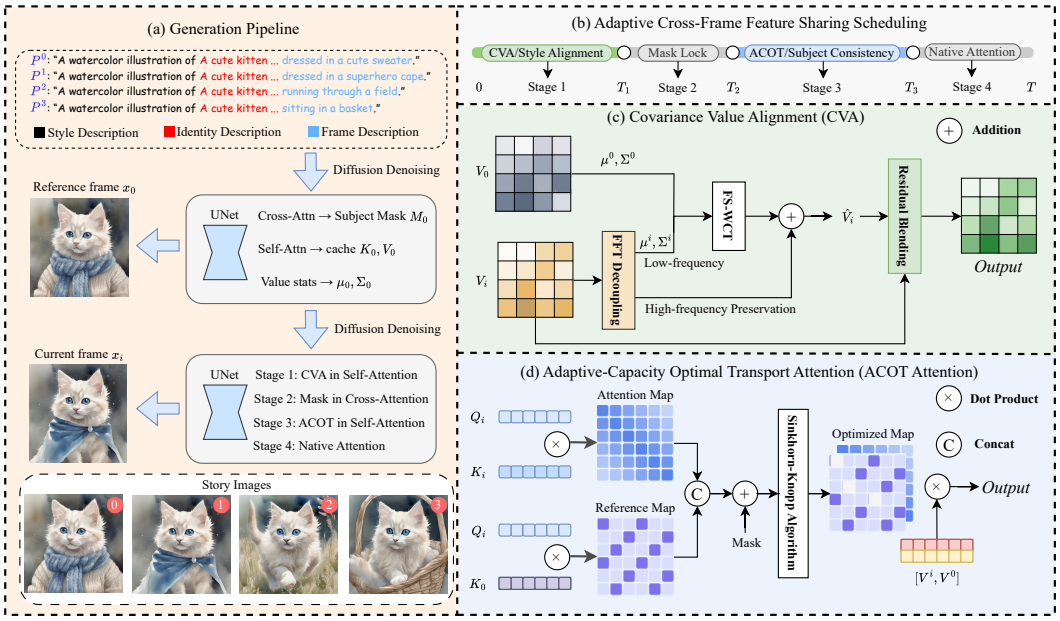


Figure 2: **(a)** The overall pipeline of AdaShare. It retains K_0, V_0, and the subject mask from the reference frame, and applies CVA and *ACOT Attention* in subsequent images to maintain style and subject consistency. **(b)** The execution schedule of CVA, subject mask extraction, and *ACOT Attention* across denoising steps. **(c)** CVA achieves style consistency via frequency-domain decomposition. **(d)** *ACOT Attention* maintains subject consistency by optimizing attention maps.

Training-Based Consistent T2I Generation. Training-based methods for consistent T2I generation typically improve identity preservation by learning subject-specific representations, fine-tuning generative model parameters, or introducing additional trainable conditioning modules. One line of work (Gal et al., 2023; Ruiz et al., 2023; Kumari et al., 2023; Tewel et al., 2023) learns personalized text embeddings or fine-tunes model parameters using reference images to incorporate target subjects into a pretrained generative model. Another line of work (Wei et al., 2023; Li et al., 2023; Ma et al., 2024; Ye et al., 2023; Li et al., 2024) trains image-conditioned encoders or adaptation modules to explicitly incorporate identity features from reference images into the generation process. These methods generally achieve strong subject consistency, but often require additional data, model training, or subject-specific optimization, which increases computational cost and limits their flexibility across different subjects and scenes.

Training-Free Consistent T2I Generation. Training-free methods (Gao et al., 2026; Park et al., 2026) directly exploit feature representations from pretrained generative models and establish cross-image feature interactions during inference. For example, ConsiStory (Tewel et al., 2024) introduces *Extended Attention* to share attention features between reference and generated images, and further improves subject consistency through feature injection. 1Prompt1Story (Liu et al., 2025) exploits linguistic context consistency and identity-preserving cross-attention to preserve subject identity across diverse scene descriptions. A recent method (Park et al., 2026) extends consistent T2I generation by considering both subject identity and visual style consistency. Despite improving cross-image consistency, existing methods still primarily rely on attention manipulation or feature sharing, while lacking explicit mechanisms to regulate reference feature distributions, and preserve high-frequency representations during style alignment.

3 METHOD

We first revisit the limitations of existing consistent text-to-image generation methods in Sec. 3.1. Specifically, subject mask extraction is often affected by noisy responses, while conventional *Extended Attention* lacks explicit constraints on reference tokens, which can degrade the quality of the generated images. Based on this analysis, we introduce *Adaptive Mask Extraction* and *Adaptive-*



Figure 3: **(Left)** Images generated using the masks extracted by the baseline method, where the subject's hand disappears in the fifth generated image. Our method better preserves this visual detail. **(Right)** The attention map and mask extracted by the baseline are more diffuse and contain substantial noise, respectively, whereas our method yields a more concentrated attention map and a cleaner mask.

Capacity Optimal Transport Attention (ACOT Attention) in Sec. 3.2. Beyond subject identity preservation, we further analyze the effect of directly applying WCT to value features in Sec. 3.1. Based on this observation, we propose *Covariance Value Alignment* (CVA) with FS-WCT in Sec. 3.3 to improve visual style consistency across generated images. Fig. 2 illustrates the overall pipeline of AdaShare.

3.1 Problem Formulation

ConsiStory (Tewel et al., 2024) employs *Extended Attention* to preserve subject consistency. The corresponding attention computation can be formulated as

$$K_{\text{ext}} = [K_i; \gamma K_0], \quad V_{\text{ext}} = [V_i; V_0],$$
$$A_{\text{ext}} = \text{Softmax}\left(\frac{Q_i K_{\text{ext}}^{\top}}{\sqrt{d}} + M\right), \tag{1}$$

where Q_i, K_i, and V_i denote the query, key, and value features of the i-th frame in the current self-attention layer, respectively. K_0 and V_0 denote the corresponding key and value features from the reference frame. γ controls the affinity strength. The operator $[\cdot;\cdot]$ denotes concatenation, d represents the dimension of each attention head, and M denotes the cross-frame visibility mask.

Definition. A_{ext} can be decomposed into $[A_{\text{self}}, A_{\text{ref}}]$, where A_{self} represents the attention map between current-frame queries Q_i and current-frame keys K_i, referred to as the Intra-Frame Attention Map (IFAMap). A_{ref} represents the attention map between current-frame queries Q_i and reference-frame keys K_0, referred to as the Reference Attention Map (RAMap). We further define the column-wise sums of IFAMap and RAMap as the Intra-Frame Attention Mass Vector (IFAMV) and the Reference Attention Mass Vector (RAMV), respectively. Specifically, IFAMV_j and RAMV_j measure the accumulated attention weight assigned to the j-th token of V_i and V_0, respectively, in the output of the current attention layer.

Problem 1. RAMV Over-Concentration. In standard *Extended Attention*, attention weights are normalized using a row-wise softmax operation. This operation independently normalizes the attention distribution for each query token in Q_i, but does not impose constraints on the global allocation of attention across reference key tokens in K_0. Consequently, if specific reference keys exhibit high feature similarity to subject-related queries in the current image, these queries may simultaneously assign excessive attention weights to a set of keys. This causes RAMV to become overly concentrated, as shown in the middle of Fig. 4, leading to imbalanced transfer of identity-related information from the reference image. As illustrated on the left of Fig. 4, images generated with *Extended Attention* exhibit incomplete subject structures, such as missing ears.

Problem 2. High-Frequency Deviation in Value Alignment. Beyond subject identity consistency, images generated within the same set should also maintain consistent visual style (Park et al., 2026). Since the value features V in self-attention encode style-related information, a straightforward strategy is to align the statistical distribution of V_i with that of V_0 using WCT. Standard WCT can be

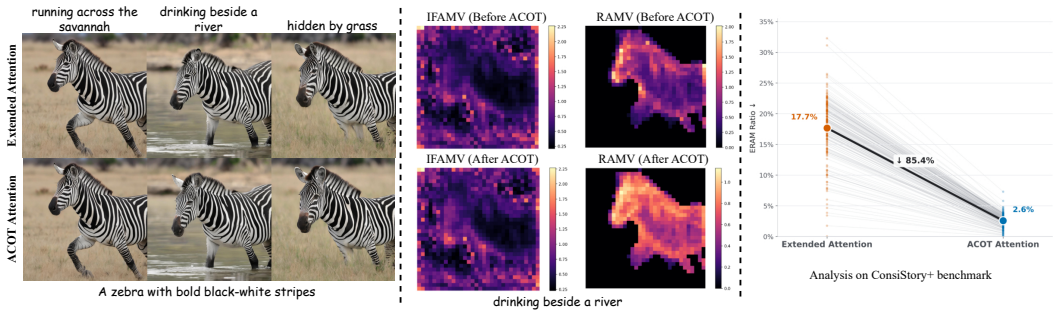


Figure 4: **(Left)** Multi-frame images generated with *Extended Attention* and our *ACOT Attention*. **(Middle)** Visualization of IFAMV and RAMV at a representative UNet layer. Our *ACOT Attention* alleviates excessive concentration on reference tokens. **(Right)** The proportion of ERAM measured during generation on the ConsiStory+ benchmark, further indicating that *ACOT Attention* constrains the distribution of RAMV.

formulated as

$$\hat{V}_i = (V_i - \mu_i)\Sigma_i^{-1/2}\Sigma_0^{1/2} + \mu_0, \tag{2}$$

where μ_i and Σ_i denote the mean and covariance statistics of the value features from the current frame, while μ_0 and Σ_0 denote the corresponding statistics from the reference frame. Although standard WCT can transfer style information from the reference frame, it also perturbs the high-frequency components of the value features, which potentially affect generation quality, as shown on the left of Fig. 5. To further analyze this effect, we measure the spectral deviation between the WCT-transformed value features and the original V_i across different radial frequency bands on the ConsiStory+ benchmark. As shown on the right of Fig. 5, standard WCT introduces notably larger spectral deviations in high-frequency regions, whereas FS-WCT better preserves the original high-frequency representations.

3.2 Adaptive Cross-Frame Sharing

3.2.1 Adaptive Mask Extraction

Previous methods (Gao et al., 2026; Tewel et al., 2024) typically estimate the subject masks by aggregating subject-related attention responses from multiple cross-attention layers. Given a denoising step t with L candidate cross-attention layers, the conventional aggregation is formulated as

$$\bar{A}_t = \frac{1}{L}\sum_l A_t^l, \tag{3}$$

where A_t^l denotes the subject-related spatial attention map from the l-th cross-attention layer at denoising step t. Since some layers produce spatially diffuse responses, directly averaging attention maps with equal weights may introduce inaccurate localization into the estimated subject mask, as shown in the right of Fig. 3. Therefore, we propose *Adaptive Mask Extraction* to obtain reliable masks. For the l-th attention layer at denoising step t, we first normalize the response of the subject token into a spatial probability distribution:

$$P_t^l(x,y) = \frac{A_t^l(x,y)}{\sum_{x,y} A_t^l(x,y) + \epsilon}, \tag{4}$$

where ϵ is a small constant for numerical stability. We then compute the spatial dispersion of the attention distribution as

$$D_t^l = \sum_{x,y} P_t^l(x,y)\left[(x-\mu_x)^2 + (y-\mu_y)^2\right], \tag{5}$$

where (μ_x, μ_y) denotes the spatial center of the attention distribution, computed as $\mu_x = \sum_{x,y} xP_t^l(x,y)$ and $\mu_y = \sum_{x,y} yP_t^l(x,y)$. A larger D_t^l indicates a more spatially diffuse attention distribution. Thus, we retain only the candidate layers satisfying $D_t^l < \tau_s$. For the 32×32 grid, the theoretical spatial variance of a uniform distribution is 170.5, with the derivation and analysis in Appendix B.1. Therefore, we set $\tau_s = 170$ to filter out diffuse responses.

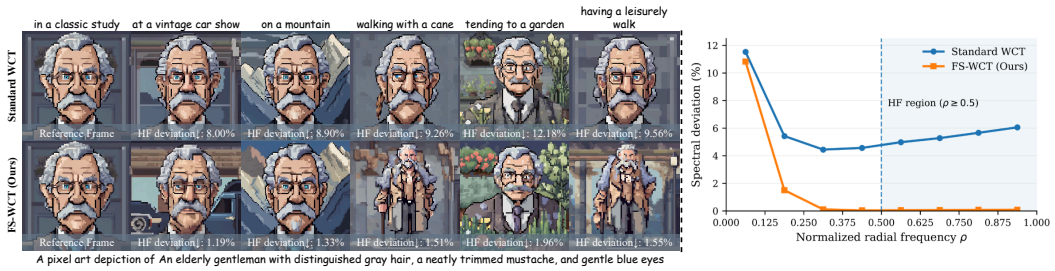


Figure 5: **(Left)** Comparison of images generated using standard WCT and our FS-WCT. **(Right)** High-frequency (HF) deviation of the radial spectrum relative to the original V on the ConsiStory+ benchmark. The radial spectral analysis further indicates that FS-WCT reduces the spectral distortion introduced by WCT, particularly in the high-frequency region ($\rho \geq 0.5$).

3.2.2 Adaptive-Capacity Optimal Transport Attention

To address **Problem 1** in Sec. 3.1, we reformulate *Extended Attention* as an optimal transport problem. Given the current-frame Q_i and K_i together with the reference-frame keys K_0, the *Extended Attention* matrix A_{ext} can be interpreted as a transport plan P (Cuturi, 2013; Litman & Guo, 2026). Standard softmax imposes only a row-wise normalization constraint, $\sum_j P_{ij} = 1$. We further introduce inequality constraints on the reference columns in A_{ref} to limit the maximum capacity of each reference token. The capacity-constrained entropy-regularized OT problem is formulated as:

$$\begin{aligned} &\min_{P \geq 0} \sum_{i=1}^{T}\sum_{j=1}^{2T} P_{ij}C_{ij} + \epsilon \sum_{i=1}^{T}\sum_{j=1}^{2T} P_{ij}\log P_{ij}, \\ &\text{s.t.} \quad \sum_{j=1}^{2T} P_{ij} = 1, \forall\, i, \sum_{i=1}^{T} P_{i,T+k} \leq \tau, \forall\, k \in \{1,\ldots,T\}. \end{aligned} \tag{6}$$

Here, the cost matrix is defined as $C = -[\frac{Q_i K_i^\top}{\sqrt{d}}, \gamma\frac{Q_i K_0^\top}{\sqrt{d}}] \in \mathbb{R}^{T\times 2T}$, where T denotes the number of tokens. ϵ denotes the entropy regularization coefficient, and τ specifies the maximum allowable attention mass for each reference token in RAMV. The transport support is restricted by the joint visibility mask M. We determine τ based on the ratio between the subject regions of the current and reference frames. According to the first-order Karush-Kuhn-Tucker (KKT) optimality conditions, the optimal transport plan can be expressed as: $P = \text{diag}(u)\mathcal{K}\,\text{diag}([\mathbf{1}_T; v])$, where $\mathcal{K} = \exp(-C/\epsilon)$ is the masked exponential kernel. From the KKT complementary slackness condition, $\beta_j(\sum_{i=1}^{T} P_{i,T+k} - \tau) = 0, \beta_k \geq 0$, the column scaling factor $v_k = \exp(-\beta_k/\epsilon)$ has the following closed-form update:

$$v_k = \min\left(1, \frac{\tau}{\hat{c}_k}\right), \qquad \text{where } \hat{c}_k = \sum_i u_i \mathcal{K}_{i,T+k}. \tag{7}$$

The complete derivation of the Lagrangian and the corresponding KKT conditions is provided in Appendix B.2. During Sinkhorn iterations, this update can be implemented by rescaling the reference columns of the attention matrix:

$$P^{(l)}_{:,T+k} \leftarrow \frac{P^{(l)}_{:,T+k}}{f_k}, \quad \text{where } f_k = \frac{1}{v_k} = \max\left(1, \frac{\hat{c}_k}{\tau}\right). \tag{8}$$

The complete iterative procedure is summarized in Algorithm 1 of Appendix B.2. We further define the attention mass in RAMV exceeding τ as Excess Reference Attention Mass (ERAM). As shown on the right of Fig. 4, *ACOT Attention* reduces the proportion of ERAM by 85.4% on the ConsiStory+ benchmark, demonstrating reduced attention concentration on reference tokens.

3.3 Covariance Value Alignment

For **Problem 2** formulated in Sec. 3.1, we observe that directly applying WCT to V may perturb high-frequency representations. To address this issue, we propose *Frequency-Selective WCT* (FS-WCT) as the core component of CVA. We first reshape the 1D sequence value features V_i of the



Figure 6: **Qualitative comparison** with state-of-the-art methods. While IP-Adapter and CoDi preserve subject consistency, they exhibit limitations in prompt fidelity and style consistency.

current frame into a 2D spatial feature map $\mathcal{V}_i$. To smoothly decompose different frequency components, we construct a centered two-dimensional Gaussian low-pass mask $M \in [0,1]^{H\times W}$:

$$M(u,v) = \exp\left(-\frac{u^2+v^2}{2r_c^2}\right), \quad u,v \in [-1,1], \tag{9}$$

where r_c controls the frequency bandwidth of the Gaussian filter, and u and v denote the normalized horizontal and vertical frequency coordinates with respect to the zero-frequency component, respectively. We then transform $\mathcal{V}_i$ into the frequency domain using a 2D Fast Fourier Transform and extract low- and high-frequency components using complementary masks:

$$V_i^{\text{low}} = \mathcal{F}_{\text{2D}}^{-1}\left(\mathcal{F}_{\text{2D}}(\mathcal{V}_i) \odot M\right), \quad V_i^{\text{high}} = \mathcal{F}_{\text{2D}}^{-1}\left(\mathcal{F}_{\text{2D}}(\mathcal{V}_i) \odot (1-M)\right), \tag{10}$$

where V_i^{low} and V_i^{high} denote the corresponding low- and high-frequency components, respectively. FS-WCT then computes the mean μ and covariance matrix Σ from the reference-frame feature V_0 and the current-frame feature V_i, and transforms the low-frequency component V_i^{low}:

$$\tilde{V}_i^{\text{low}} = (V_i^{\text{low}} - \mu_i)\Sigma_i^{-1/2}\Sigma_0^{1/2} + \mu_0. \tag{11}$$

This transformation transfers the global visual style statistics of the reference frame to the low-frequency representation while maintaining the original high-frequency information. We introduce a blending coefficient $\alpha \in [0,1]$ to control the strength of reference style alignment:

$$V_i^{\text{out}} = \alpha\left(\tilde{V}_i^{\text{low}} + V_i^{\text{high}}\right) + (1-\alpha)V_i. \tag{12}$$

Table 1: **Quantitative comparison.** The best and second-best results are highlighted in **bold** and underlined, respectively. Vanilla SDXL is shown as a reference and excluded from the comparison.

Methods	Base Model	Training-Free	DINO↑	VGG-Style↓	CLIP-T↑	CLIP-I↑	DreamSim↓	FID↓
Vanilla SDXL	–	–	0.6060	4.3890	0.9136	0.8775	0.3364	–
UNO	FLUX	×	0.7641	3.2777	0.8519	0.9131	0.2246	–
PhotoMaker	SDXL	×	0.5866	8.6452	0.8895	0.8679	0.3757	54.7263
IP-Adapter	SDXL	×	0.7963	1.9521	0.8555	0.9293	0.2065	60.5017
ConsiStory	SDXL	✓	0.7068	4.8126	**0.9160**	0.9005	0.2561	53.3715
StoryDiffusion	SDXL	✓	0.7792	3.2639	0.8966	0.9123	0.2276	63.3608
1Prompt1Story	SDXL	✓	0.7660	1.2757	0.8905	0.9189	0.2043	57.4361
CoDi	SDXL	✓	0.7198	6.5864	0.8997	0.8967	0.2723	55.9575
Infinite-Story	Infinity	✓	**0.8140**	2.1085	0.8778	0.9233	0.1808	–
Ours	SDXL	✓	0.7963	**1.0588**	0.9012	**0.9298**	**0.1692**	**45.7914**

Finally, the aligned value features V_i^{out} are fed into the self-attention layer.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUPS

Benchmark and Comparison Methods. We evaluate AdaShare on the ConsiStory+ benchmark (Liu et al., 2025) for consistent text-to-image generation. ConsiStory+ contains 200 prompt sets, where each set consists of 5–10 prompts, resulting in over 1,100 generated images in total. We compare AdaShare with representative baseline methods, including UNO (Wu et al., 2025), PhotoMaker (Li et al., 2024), IP-Adapter (Ye et al., 2023), ConsiStory (Tewel et al., 2024), StoryDiffusion (Zhou et al., 2024), 1Prompt1Story (Liu et al., 2025), CoDi (Gao et al., 2026), and Infinite-Story (Park et al., 2026). All baseline methods are reproduced using their officially released implementations. AdaShare is implemented based on SDXL. Detailed model configurations and additional results are provided in the Appendix A and Appendix D.

Evaluation Metrics. We evaluate prompt fidelity using CLIP-T (Hessel et al., 2021). For subject consistency, we adopt CLIP-I (Hessel et al., 2021) and DreamSim (Fu et al., 2023). Following the DreamSim evaluation protocol, we remove backgrounds using CarveKit and replace them with random noise to reduce background bias. Style consistency is evaluated using DINO (Caron et al., 2021) and VGG-Style (Gatys et al., 2016) metrics. To evaluate overall generation quality, we compute the FID (Heusel et al., 2017) with respect to images generated by the base SDXL model.

4.2 EXPERIMENTAL RESULTS

Qualitative Comparison. Fig. 6 presents qualitative comparisons with the baselines. AdaShare achieves a favorable trade-off among text-image alignment, subject consistency, and style consistency. In contrast, although Infinite-Story achieves competitive consistency, it occasionally produces visual artifacts in certain cases, such as duplicated tails in the fox example. Additionally, CoDi and 1Prompt1Story exhibit limitations in prompt fidelity and style consistency. Notably, AdaShare can be applied to DiT-based models (Black Forest Labs, 2024), such as FLUX. The implementation details and qualitative results are provided in Appendix A.2 and Appendix D.4, respectively.

Quantitative Comparison. Table 1 reports quantitative comparison results. AdaShare achieves competitive performance across subject consistency, style consistency, and generation quality. Specifically, our method yields the highest CLIP-I and lowest DreamSim scores, demonstrating superior subject identity preservation during cross-image generation. Furthermore, AdaShare achieves the best VGG-Style score and a competitive DINO score (0.7963), slightly lower than the Infinity-based Infinite-Story. The higher DINO score of Infinite-Story may be related to the structural constraints introduced by its Infinity backbone. In comparison, AdaShare is based on SDXL, which provides more flexibility for prompt-driven generation. Meanwhile, VGG-Style prioritizes low-level color distributions and local stylistic textures, serving as a more representative proxy for evaluating artistic style preservation. More importantly, AdaShare maintains a competitive CLIP-T score and the lowest FID, suggesting that consistency does not substantially degrade text-image alignment and image quality.

Table 2: **User study** conducted with 32 participants evaluating human preference across text-to-image generation methods.

Method	1Prompt1Story	Infinite-Story	CoDi	Ours
Percent (%) ↑	21.1	21.3	18.6	**39.0**

Table 3: **Ablation studies** on model components and mask strategy. Results demonstrate the effectiveness of CVA, *ACOT Attention*, and *Adaptive Mask Extraction* in balancing consistency, prompt fidelity, and generation quality.

Configuration	VGG-Style ↓	CLIP-T ↑	CLIP-I ↑	DreamSim ↓	FID ↓
(a) Architectural Components					
w/ CVA	1.3170	0.9035	0.9135	0.2220	43.3768
w/ ACOT Attn.	2.3955	**0.9083**	0.9084	0.2304	**42.4033**
(b) Mask Strategy					
w/ Baseline Mask	1.0649	0.9021	0.9288	0.1728	47.7851
Full Model (Ours)	**1.0588**	0.9012	**0.9298**	**0.1692**	45.7914



Figure 7: **Qualitative ablation studies** on architectural components and mask strategy.

User Study. To evaluate human preference for generated images in terms of text-image alignment, subject consistency, image quality, and style consistency, we conduct a user study comparing AdaShare with 1Prompt1Story, Infinite-Story, and CoDi. Following prior work, we randomly select 30 prompt sets from the ConsiStory+ benchmark, each containing four fixed-length prompts for image generation. A total of 32 participants are asked to select the method that achieves the best overall performance. As reported in Table 2, our method receives the highest human preference rate. Additional details of the user study are provided in Appendix E.

Ablation study. To evaluate each component and the subject mask extraction strategy, quantitative and qualitative results are in Table 3 and Fig. 7, respectively. In Table 3, employing only *ACOT Attention* yields the highest CLIP-T and lowest FID, showing it preserves textual semantics and image fidelity. Using only CVA reduces the VGG-Style distance to 1.3170, promoting cross-image style consistency. Furthermore, using only the Baseline Mask degrades VGG-Style, CLIP-I, DreamSim, and FID, showing that accurate subject localization is critical for consistent feature sharing. The full model performs best on VGG-Style, CLIP-I, and DreamSim, demonstrating the complementary effects of the three components. Additional ablations are in Appendix C.

In Fig. 7, *ACOT Attention* alone improves subject alignment with scene descriptions, while background color style varies. CVA alone yields a more consistent watercolor style, yet subject appearance varies across scenes. The Baseline Mask alone causes local appearance variations under complex poses. In contrast, the full model remains consistent across all scenes, confirming complementary benefits of the three components for subject and style consistency and feature sharing.

5 CONCLUSION

In this paper, we proposed AdaShare, a training-free framework that addresses incomplete subject generation and degraded image quality caused by unreliable cross-frame feature sharing. *Adaptive Mask Extraction* refines subject localization and activates feature sharing after subject regions stabilize. *ACOT Attention* formulates *Extended Attention* as an optimal transport problem with adaptive capacity constraints, alleviating excessive attention concentration on reference tokens caused by token imbalance. *Covariance Value Alignment* aligns low-frequency value statistics with the reference while preserving high-frequency representations, promoting style consistency. Extensive experiments on ConsiStory+ demonstrate that AdaShare achieves competitive performance against training-free baselines across consistency metrics and human evaluations while maintaining prompt fidelity and generation quality. These findings support adaptive cross-frame feature sharing as an effective approach to improving subject and style consistency without additional training.

AI USE STATEMENT

We used LLMs to improve the wording of the manuscript and to help find related papers. We checked suggested references against the original publications and reviewed all edits before submission. The authors developed the research questions, methods, and experiments, interpreted the results, and take responsibility for the final paper.

REFERENCES

Black Forest Labs. FLUX.1 [dev]. https://huggingface.co/black-forest-labs/FLUX.1-dev, 2024. Accessed: 2026-09-24.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. SAM 3: Segment anything with concepts. In *International Conference on Learning Representations (ICLR)*, 2026.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *International Conference on Computer Vision (ICCV)*, pp. 9630–9640, 2021.

Tai-Yin Chiu. Understanding generalized whitening and coloring transform for universal style transfer. In *International Conference on Computer Vision (ICCV)*, pp. 4451–4459, 2019.

Jooyoung Choi, Chaehun Shin, Yeongtak Oh, Heeseung Kim, Jungbeom Lee, and Sungroh Yoon. Style-friendly SNR sampler for style-driven generation. In *Winter Conference on Applications of Computer Vision (WACV)*, pp. 5703–5713, 2026.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 26, 2013.

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 *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 50742–50768, 2023.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *International Conference on Learning Representations (ICLR)*, 2023.

Zhanxin Gao, Beier Zhu, Jian Yang, and Ying Tai. CoDi: Subject-consistent and pose-diverse text-to-image generation. In *International Conference on Learning Representations (ICLR)*, 2026.

Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 2414–2423, 2016.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. In *International Conference on Learning Representations (ICLR)*, 2023.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7514–7528, 2021.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 6840–6851, 2020.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 36652–36663, 2023.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 1931–1941, 2023.

Dongxu Li, Junnan Li, and Steven Hoi. BLIP-Diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 30146–30166, 2023.

Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. PhotoMaker: Customizing realistic human photos via stacked ID embedding. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 8640–8650, 2024.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision (ECCV)*, pp. 366–384. Springer, 2024.

Elon Litman and Gabe Guo. You need better attention priors. In *International Conference on Machine Learning (ICML)*, 2026.

Tao Liu, Kai Wang, Senmao Li, Joost Van de Weijer, Fahad Khan, Shiqi Yang, Yaxing Wang, Jian Yang, and Ming-Ming Cheng. One-prompt-one-story: Free-lunch consistent text-to-image generation using a single prompt. In *International Conference on Learning Representations (ICLR)*, 2025.

Jian Ma, Junhao Liang, Chen Chen, and Haonan Lu. Subject-diffusion: Open domain personalized text-to-image generation without test-time fine-tuning. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–12, 2024.

Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning (ICML)*, 2022.

Jihun Park, Kyoungmin Lee, Jongmin Gim, Hyeonseo Jo, Minseok Oh, Wonhyeok Choi, Kyumin Hwang, Jaeyeul Kim, Minwoo Choi, and Sunghoon Im. Infinite-story: A training-free consistent text-to-image generation. In *AAAI Conference on Artificial Intelligence*, volume 40, pp. 8278–8286, 2026.

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 *International Conference on Learning Representations (ICLR)*, 2024.

Jiang Qin, Alexandra Gomez-Villa, Senmao Li, Shiqi Yang, Yaxing Wang, Kai Wang, and Joost van de Weijer. Free-lunch color-texture disentanglement for stylized image generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

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

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 10674–10685, 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 *Computer Vision and Pattern Recognition (CVPR)*, pp. 22500–22510, 2023.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 36479–36494, 2022.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 25278–25294, 2022.

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

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

Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. ELITE: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *International Conference on Computer Vision (ICCV)*, pp. 15897–15907, 2023.

Shaojin Wu, Mengqi Huang, Wenxu Wu, Yufeng Cheng, Fei Ding, and Qian He. Less-to-more generalization: Unlocking more controllability by in-context generation. In *International Conference on Computer Vision (ICCV)*, pp. 18682–18692, 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.

Sixian Zhang, Bohan Wang, Junqiang Wu, Yan Li, Tingting Gao, Di Zhang, and Zhongyuan Wang. Learning multi-dimensional human preference for text-to-image generation. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 8018–8027, 2024.

Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. StoryDiffusion: Consistent self-attention for long-range image and video generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 110315–110340, 2024.

APPENDIX

A IMPLEMENTATION DETAILS

A.1 ADAPTIVE CROSS-FRAME FEATURE SHARING SCHEDULING

A fixed cross-frame feature-sharing schedule may not adequately accommodate the varying generation dynamics across subjects and scenes. To address this limitation, we introduce a state-driven adaptive scheduling strategy that dynamically transitions among CVA, *ACOT Attention*, and Native Attention based on subject-mask stabilization and the convergence of cross-frame feature evolution.

For SDXL-based methods, style information begins to emerge during the early denoising stages (Choi et al., 2026). Accordingly, we first perform style alignment for the non-reference frames. From denoising steps $T_1 = 10$ to $T_2 = 15$, we progressively identify the subject mask and compute the IoU between masks obtained at consecutive steps. Furthermore, we binarize the aggregated map using Otsu's threshold and retain the largest connected component as the subject mask. *Adaptive Mask Extraction* extracts subject-specific cross-attention maps from the bottleneck and the first decoder block of the UNet. Once the IoU reaches 0.85, or when denoising step $T_2 = 15$ is reached, the current subject mask is fixed, after which the generation process transitions to the subsequent subject consistency alignment stage.

During subject alignment with *ACOT Attention*, we assess the stabilization of cross-frame alignment using two complementary signals that capture the temporal variation of reference attention and UNet feature evolution. Specifically, at each denoising step t, we compute the entropy of the reference attention at each ACOT layer and average these values across layers to obtain H_t. We then define $\Delta H_t = |H_t - H_{t-1}|$, which measures the step-to-step variation in reference attention entropy. A small ΔH_t indicates that the subject-level correspondence between the current and reference frames has become temporally stable. In parallel, we characterize UNet feature evolution using the normalized feature displacement between consecutive denoising steps,

$$E_t = \frac{\|F_t - F_{t-1}\|_2}{\|F_{t-1}\|_2 + \epsilon}, \tag{13}$$

where F_t denotes the UNet feature representation at denoising step t. Since E_t captures the magnitude of feature evolution rather than its temporal stability, we further measure its step-to-step variation as $\Delta E_t = |E_t - E_{t-1}|$. A small ΔE_t indicates that the magnitude of feature evolution varies only slightly across successive denoising steps.

We combine these two signals to define a joint stabilization criterion. After a minimum of three ACOT steps, we regard cross-frame alignment as sufficiently stabilized when both $\Delta H_t < 0.025$ and $\Delta E_t < 0.005$, and terminate *ACOT Attention* early. If either criterion is not satisfied, *ACOT Attention* remains active until the predefined maximum denoising step. Native Attention is then restored for the remaining denoising steps, allowing subsequent refinement to proceed without unnecessary cross-frame constraints after subject alignment has stabilized.

A.2 FLUX-BASED IMPLEMENTATION

For the FLUX-based implementation, we extend our method from SDXL to FLUX.1-dev (Black Forest Labs, 2024) to evaluate its applicability to a different generative architecture. For subject-mask extraction, we leverage the joint-attention maps in the double-stream Transformer blocks of FLUX, where text and image queries and keys participate in a shared attention operation. We therefore extract subject-related attention responses from both the image-to-text and text-to-image directions and fuse them to obtain a more robust subject mask.

For *ACOT Attention*, we retain the native FLUX joint-attention formulation and augment its key and value sequences with the reference-frame keys K_0 and values V_0, respectively. The capacity constraint is applied exclusively to the newly introduced reference-image attention columns, without directly modifying the native text-image attention entries. Additional FLUX-based generation results are provided in Appendix D.4.



Figure 8: **Qualitative comparison of subject masks** derived from cross-attention maps. The subject masks segmented by SAM3 are used as references. The results indicate that our method captures most of the subject regions.

Table 4: **Quantitative comparison of subject masks** derived from cross-attention maps. Time denotes the elapsed time from diffusion step 0 to mask extraction. All runtime measurements are conducted on an NVIDIA RTX 3090 GPU.

Method	Time (s)	IoU (%) ↑
CoDi	19.57	61.87
Baseline	3.32	61.07
Ours	3.86	**62.22**

A.3 Implementation of the comparison methods

All baseline methods are reproduced using their publicly available implementations. For UNO, IP-Adapter and PhotoMaker, which require an additional reference image, we first generate a reference image using the corresponding base model, conditioned on the identity description. The generated reference image is then shared across all frames within the corresponding prompt set.

B Theoretical Analysis and Proofs

B.1 Theoretical Spatial Variance of Uniform Attention

Consider a uniform attention map defined over an $N \times N$ spatial grid, where each spatial position is assigned equal probability mass:

$$P(x, y) = \frac{1}{N^2}, \qquad x, y \in \{0, \ldots, N-1\}. \tag{14}$$

The corresponding spatial centroid is located at the center of the grid:

$$\mu_x = \mu_y = \frac{N-1}{2}. \tag{15}$$

The spatial dispersion is defined as:

$$D = \sum_{x,y} P(x, y) \left[(x - \mu_x)^2 + (y - \mu_y)^2 \right]. \tag{16}$$

Algorithm 1 Capacity-Sinkhorn Algorithm for *ACOT Attention*

Require: Initial transport matrix $P^{(0)}$, reference capacity τ, iteration number L, the current frame V_i and the reference frame V_0
Ensure: *ACOT attention* output Output
1: **for** $l = 1, \ldots, L$ **do**
2: Update the row scaling factor $u^{(l)}$ according to Eq. 26.
3: Perform row normalization of $P^{(l-1)}$ using $u^{(l)}$.
4: Compute the reference token mass $\hat{c}^{(l)}$ according to Eq. 27.
5: Update the capacity scaling factor $v^{(l)}$ according to Eq. 29.
6: Update the transport matrix $P^{(l)}$ using the capacity-adjusted reference transport.
7: **end for**
8: Obtain the final transport matrix $P^* = P^{(L)}$.
9: Normalize P^* along the key dimension to obtain the attention matrix $\hat{P}$.
10: Compute *ACOT attention* output:Output $= \hat{P}[V_i; V_0]$
11: **return** Output

Since the two spatial dimensions follow identical discrete uniform distributions, D can be decomposed into the sum of their marginal variances:

$$D_{\text{uniform}} = \text{Var}(x) + \text{Var}(y). \tag{17}$$

For a one-dimensional discrete uniform distribution over $0, \ldots, N-1$, the variance is

$$\text{Var}(x) = \frac{1}{N} \sum_{x=0}^{N-1} \left(x - \frac{N-1}{2} \right)^2 = \frac{N^2 - 1}{12}. \tag{18}$$

Therefore, the spatial dispersion of an $N \times N$ uniform attention map is

$$D_{\text{uniform}} = 2 \cdot \frac{N^2 - 1}{12} = \frac{N^2 - 1}{6}. \tag{19}$$

For $N = 32$, this yields $D_{\text{uniform}} = 170.5$, which serves as the reference value for spatial dispersion. Accordingly, we set $\tau_s = 170$, slightly below the theoretical uniform-attention dispersion of 170.5, to filter out highly diffuse attention maps that approach the uniform case while retaining spatially localized responses for subsequent subject mask construction.

To evaluate the effectiveness of our mask extraction strategy, Fig. 8 qualitatively compares the subject masks produced by our method with those obtained from the baseline and CoDi. Specifically, we use the masks generated by SAM3 (Carion et al., 2026) as reference masks for evaluation. Compared with the alternative methods, our approach suppresses most background responses while retaining the majority of subject regions. For quantitative evaluation, Table 4 reports the IoU between the masks produced by each method and the corresponding SAM3 reference masks, along with the mask extraction time. Our method achieves the highest IoU among the compared approaches, suggesting that spatial dispersion filtering improves the agreement between the extracted masks and the SAM3 references. In addition, CoDi obtains the subject mask at the final diffusion step and therefore requires the full denoising process to be completed before mask extraction. In contrast, our method extracts the subject mask at an earlier denoising stage, thereby achieving lower extraction latency.

B.2 Derivation of Adaptive-Capacity Optimal Transport Attention

We derive *Adaptive-Capacity Optimal Transport Attention* (*ACOT Attention*) from the constrained optimization problem introduced in the main text. We use $j \in 1, \ldots, 2T$ to index all columns of the transport plan and $k \in 1, \ldots, T$ to index the reference columns. The transport support is restricted by the joint visibility mask M.

The transport plan $P \in \mathbb{R}_+^{T\times 2T}$ is obtained by solving the following constrained optimization problem:

$$\begin{aligned} \min_{P\geq 0} \quad & \sum_{i=1}^{T}\sum_{j=1}^{2T} P_{ij}C_{ij} + \epsilon \sum_{i=1}^{T}\sum_{j=1}^{2T} P_{ij}\log P_{ij} \\ \text{s.t.} \quad & \sum_{j=1}^{2T} P_{ij} = 1, \qquad \forall i, \\ & \sum_{i=1}^{T} P_{i,T+k} \leq \tau, \qquad \forall k. \end{aligned} \tag{20}$$

The first constraint maintains row-wise attention normalization, whereas the second restricts the total attention weight allocated to each reference token. Additionally, M represents the cross-frame visibility mask.

Introducing dual variables $\alpha_i \in \mathbb{R}$ and $\beta_k \geq 0$ for the row-normalization and reference-capacity constraints, respectively, gives

$$\begin{aligned} \mathcal{L} = & \sum_{i,j} (P_{ij}C_{ij} + \epsilon P_{ij}\log P_{ij}) + \sum_i \alpha_i \left(\sum_j P_{ij} - 1\right) \\ & + \sum_k \beta_k \left(\sum_i P_{i,T+k} - \tau\right). \end{aligned} \tag{21}$$

The KKT stationarity conditions give

$$P_{ij} = u_i \mathcal{K}_{ij}, \qquad j \leq T, \tag{22}$$

and

$$P_{i,T+k} = u_i \mathcal{K}_{i,T+k} v_k, \tag{23}$$

where

$$\mathcal{K} = \exp\left(-\frac{C}{\epsilon}\right), \qquad u_i = \exp\left(-\frac{\alpha_i + \epsilon}{\epsilon}\right), \qquad v_k = \exp\left(-\frac{\beta_k}{\epsilon}\right). \tag{24}$$

Since $\beta_k \geq 0$, we have $0 < v_k \leq 1$. Therefore, the transport plan can be expressed as

$$\boxed{P = \mathrm{diag}(u)\,\mathcal{K}\,\mathrm{diag}\left(\begin{bmatrix}\mathbf{1}_T \\ v\end{bmatrix}\right)}. \tag{25}$$

This form shows that *ACOT Attention* preserves the current-frame columns while selectively suppressing reference columns through v.

From the row-normalization constraint, the row-scaling factor is

$$\boxed{u_i = \left(\sum_{j=1}^{T} \mathcal{K}_{ij} + \sum_{k=1}^{T} \mathcal{K}_{i,T+k} v_k\right)^{-1}}. \tag{26}$$

For fixed u, define the reference-column mass before capacity scaling as

$$\hat{c}_k = \sum_{i=1}^{T} u_i \mathcal{K}_{i,T+k}. \tag{27}$$

The actual mass of the k-th reference column is then $c_k = v_k \hat{c}_k$. According to the complementary-slackness condition

$$\beta_k(c_k - \tau) = 0, \tag{28}$$

the capacity constraint is inactive when $\hat{c}_k \leq \tau$, yielding $v_k = 1$, and active otherwise, yielding $v_k \hat{c}_k = \tau$. Thus,

$$\boxed{v_k = \min\left(1, \frac{\tau}{\hat{c}_k}\right)}. \tag{29}$$

Table 5: **Ablation study on the hyperparameter** α (Eq. 12). The remaining hyperparameters are set to $\gamma = 1.0$ (Eq. 1), $L = 1$ (Algorithm 1), and $r_c = 0.25$ (Eq. 9) by default.

α	VGG-Style↓	CLIP-T↑	CLIP-I↑	DreamSim↓	FID↓
0.4	1.7981	0.9060	0.9142	0.2191	43.0260
0.5	1.2700	0.9042	0.9160	0.2076	43.8753
0.6	1.1318	0.9018	0.9203	0.1966	44.6836
0.7	1.0167	0.9000	0.9224	0.1890	45.4882

Table 6: **Ablation study on the hyperparameter** r_c (Eq. 9). The remaining hyperparameters are set to $\gamma = 1.25$ (Eq. 1), $L = 2$ (Algorithm 1), and $\alpha = 0.6$ (Eq. 12) by default.

r_c	VGG-Style↓	CLIP-T↑	CLIP-I↑	DreamSim↓	FID↓
0.15	1.0588	0.9012	0.9298	0.1692	45.7914
0.25	1.0309	0.9009	0.9301	0.1697	46.2409
0.50	0.9727	0.8988	0.9321	0.1631	46.9275
w/ Standard WCT	0.9025	0.8975	0.9329	0.1597	47.5041

Relation to Extended Attention. Let $C = -S$, where S denotes the *Extended Attention* logits. Then

$$\mathcal{K} = \exp(S/\epsilon). \tag{30}$$

When all reference-capacity constraints are inactive, $v_k = 1$ for all k, and *ACOT Attention* reduces to row normalization of $\mathcal{K}$:

$$P_{ij} = \frac{\exp(S_{ij}/\epsilon)}{\sum_{m=1}^{2T} \exp(S_{im}/\epsilon)}. \tag{31}$$

For $\epsilon = 1$, this becomes

$$P = \mathrm{Softmax}(S), \tag{32}$$

recovering conventional *Extended Attention*. Therefore, *ACOT Attention* can be interpreted as a capacity-constrained generalization of *Extended Attention* that prevents excessive attention accumulation on individual reference tokens.

B.3 Reference Capacity τ

We adapt the reference-token capacity according to the relative subject scale between the current and reference frames. Let $\mathcal{M}_i$ and $\mathcal{M}_0$ denote the subject masks of the current frame and the reference frame, respectively. Accordingly, we define the reference capacity τ as

$$\tau = \frac{|\mathcal{M}_i|}{|\mathcal{M}_0|}, \tag{33}$$

where $|\mathcal{M}_i|$ denotes the number of subject-related tokens in the current frame, and $|\mathcal{M}_0|$ denotes the number of reference subject tokens in the reference frame. By definition, RAMV_j denotes the total attention mass accumulated by the j-th reference token over all current-frame queries. Consequently, the ratio $|\mathcal{M}_i|/|\mathcal{M}_0|$ represents the average matching load per reference token. We use this ratio as the adaptive baseline for the reference capacity to determine whether a specific reference token receives excessive attention mass.

C Additional Ablation Study

C.1 The Ablation of Covariance Value Alignment

Table 5 reports the ablation study on the coefficient α in Eq. 12. As α increases, VGG-Style, CLIP-I, and DreamSim consistently improve, indicating that stronger feature injection promotes style and subject consistency. In contrast, the performance measured by CLIP-T, and FID gradually deteriorates, suggesting that excessive feature injection restricts generation flexibility and compromises

Table 7: **Ablation study of affinity strengths** γ (Eq. 1). The remaining hyperparameters are set to $r_c = 0.25$ (Eq. 9), $L = 1$ (Algorithm 1), and $\alpha = 0.6$ (Eq. 12) by default.

Method	γ	VGG-Style↓	CLIP-T↑	CLIP-I↑	DreamSim↓	FID↓
Extended Attention	1.00	1.1271	0.9021	0.9211	0.1955	44.8160
ACOT Attention	1.00	1.1318	0.9018	0.9203	0.1966	44.6836
Extended Attention	1.25	1.0408	0.8998	0.9318	0.1622	46.7600
ACOT Attention	1.25	1.0332	0.9005	0.9295	0.1682	46.5232
Extended Attention	1.50	1.1216	0.8967	0.9342	0.1528	48.3351
ACOT Attention	1.50	1.0724	0.8974	0.9362	0.1501	48.0342

Table 8: **Ablation study of** L (Algorithm 1). The remaining hyperparameters are set to $\gamma = 1.25$ (Eq. 1), $r_c = 0.25$ (Eq. 9), and $\alpha = 0.6$ (Eq. 12) by default.

L	VGG-Style↓	CLIP-T↑	CLIP-I↑	DreamSim↓	FID↓
1	1.0332	0.9005	0.9295	0.1682	46.5232
2	1.0309	0.9009	0.9301	0.1697	46.2409
3	1.0324	0.9009	0.9293	0.1706	46.1128
4	1.0331	0.9008	0.9294	0.1713	46.1666

overall image quality. $\alpha = 0.6$ provides a balance between these two groups of metrics, improving consistency while maintaining favorable text fidelity and generation quality. We therefore set $\alpha = 0.6$ by default.

Table 6 reports the ablation study on the frequency cutoff ratio r_c in Eq. 9. As r_c increases, the FS-WCT alignment progressively covers a broader range of frequency components. Overall, VGG-Style, CLIP-I, and DreamSim exhibit improved consistency, whereas CLIP-T decreases and FID increases. The results with standard WCT, which performs alignment over the full frequency spectrum, further reinforce this trend. Aligning statistics over the full frequency spectrum yields stronger consistency, but results in lower text alignment and degraded generation quality. In contrast, $r_c = 0.15$ restricts the alignment to lower-frequency components, while preserving high-frequency details to a greater extent. This setting provides a more favorable trade-off between consistency and generation quality. We therefore set $r_c = 0.15$ by default.

C.2 Effect of ACOT Attention

Table 7 compares *Extended Attention* and *ACOT Attention* under different affinity strengths γ (Eq. 1). Overall, the benefit of *ACOT Attention* becomes more evident as γ increases. When $\gamma = 1.0$, the attention assigned to the reference tokens is relatively weak, and the two methods achieve comparable performance. As γ increases, *Extended Attention* tends to concentrate attention mass on high-response reference tokens. In contrast, *ACOT Attention* alleviates this excessive concentration through capacity constraints. At $\gamma = 1.5$, it improves all five metrics over *Extended Attention*. This trend is consistent with the motivation of *ACOT Attention* to mitigate the excessive concentration of Reference Attention Mass. Therefore, we adopt $\gamma = 1.25$ as the default configuration, as it provides a favorable trade-off among subject consistency, style consistency, prompt fidelity, and generation quality.

Fig. 9 further provides a qualitative comparison between *ACOT Attention* and *Extended Attention*. When using *Extended Attention*, the generated subsequent frames may exhibit incomplete local structures, such as partially generated zebra heads in the zebra sequence and missing hand details in the woman sequence. In contrast, *ACOT Attention* improves prompt fidelity and image quality by reducing the excessive concentration of certain tokens in the Reference Attention Map.

Table 8 compares the performance of different iteration numbers in Algorithm 1. When the iteration number L increases from 1 to 2, VGG-Style, CLIP-T, CLIP-I, and FID are improved, suggesting that an additional iteration further optimizes the attention allocation and alleviates the excessive con-

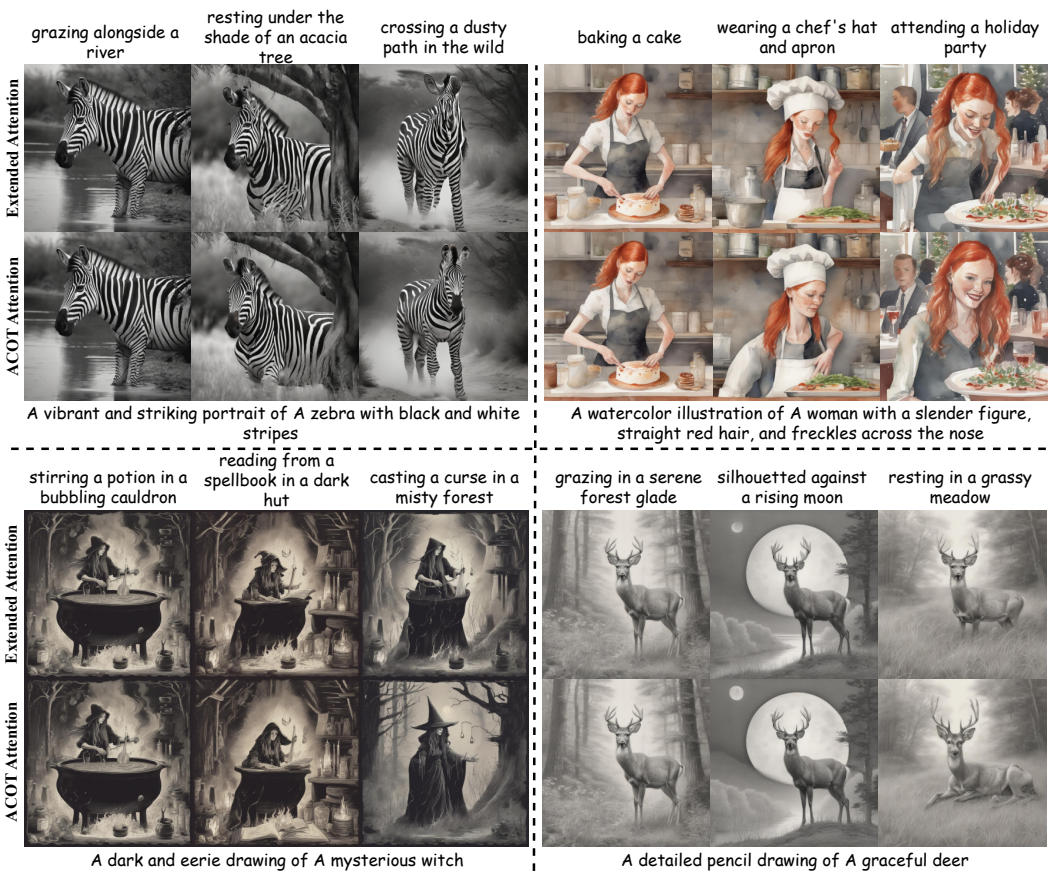


Figure 9: **Qualitative comparison between *Extended Attention* and *ACOT Attention*.** Our *ACOT Attention* generates images with higher visual quality and better text alignment.

Table 9: **Comparison on additional metrics.** The best, second-best, and third-best values are highlighted in blue, red, and green, respectively.

Metric	PhotoMaker	IP-Adapter	ConsiStory	StoryDiffusion	1Prompt1Story	CoDi	Ours
Aesthetic Score ↑	6.0260	6.0630	6.1692	6.0377	6.2899	6.1311	6.2296
VQAScore ↑	0.7289	0.6360	0.7754	0.7057	0.7201	0.7331	0.7502
PickScore ↑	22.4589	22.7918	23.4435	22.9489	23.2242	23.0653	23.2264
MPS ↑	13.9819	14.6648	15.4024	15.2197	14.8689	14.8135	15.0347

centration of reference features. Further increasing L to 3 or 4 yields only marginal improvements in FID, while VGG-Style, CLIP-I, and DreamSim exhibit varying degrees of degradation. These results suggest that the performance of *ACOT Attention* largely saturates after two iterations, with additional iterations providing limited gains while incurring extra computational overhead. Therefore, we set $L = 2$ as the default configuration for all experiments.

D ADDITIONAL RESULTS

D.1 ADDITIONAL METRICS

To provide a more comprehensive comparison across multiple evaluation dimensions, we further report Aesthetic Score (Schuhmann et al., 2022), VQAScore (Lin et al., 2024), PickScore (Kirstain et al., 2023), and MPS (Zhang et al., 2024) for SDXL-based methods on the ConsiStory+ bench-



Figure 10: **Multi-subject generation.** The results show that AdaShare maintains subject consistency and appearance consistency when generating images with multiple subjects.

Table 10: **Inference time and memory usage.** We report the inference time and peak GPU memory consumption when generating a set of five images at a resolution of 1024×1024 from the prompt set on a single RTX 4090 GPU with 48GB VRAM.

Method	Inference time (s)	GPU memory (GB)
1Prompt1Story	71.50	16.44
ConsiStory	83.88	46.48
CoDi	99.29	44.64
StoryDiffusion	37.48	40.58
Ours	85.01	42.36

mark. As shown in Table 9, although our method is primarily designed to improve subject consistency, it remains competitive across all four complementary metrics. Specifically, our method ranks second in Aesthetic Score, VQAScore, and PickScore, and third in MPS. These results suggest that the proposed consistency mechanism enhances subject consistency without substantially compromising general generation quality.

D.2 MULTI-SUBJECT GENERATION

Compared with single-subject generation, multi-subject generation mainly extends the reference token access in *Adaptive Mask Extraction* and *ACOT Attention*. First, we independently extract a mask for each subject. Once the masks for all subjects have been determined, we disable style injection and activate *ACOT Attention* for the remaining denoising steps. Second, the current and reference masks of each subject are used to restrict cross-image attention, such that each subject can attend only to the reference key/value features associated with the same subject. This subject-specific masking prevents cross-subject access to reference features. As shown in Fig. 10, our method preserves the identity of each subject while maintaining consistency across the generated images.

D.3 INFERENCE TIME AND MEMORY USAGE

As shown in Table 10, our method achieves competitive inference speed and GPU memory usage compared with existing consistency generation methods. This efficiency is consistent with our design, which avoids extensive caching of intermediate features for cross-image consistency modeling. Although the adaptive optimization process introduces additional computational overhead, our method maintains a competitive balance between computational efficiency, generation quality, and subject consistency.

D.4 Generation based on DiT architectures

To verify the applicability of our method to DiT-based models, we apply it to FLUX, and the generated results are shown in Fig. 11. The results indicate that our method can preserve subject consistency and style consistency in FLUX.

D.5 More Visualizations

As shown in Fig. 12, we provide additional generation results. These results indicate that our method maintains subject consistency and appearance consistency across a wider range of scenarios while preserving prompt fidelity.

E User Study Details

In our user study, we compare our method with recent training-free approaches, including 1Prompt1Story, Infinite-Story, and CoDi. We randomly select 30 prompt sets from the ConsiStory+ benchmark and generate four frames for each prompt set using each method.

As illustrated in Fig. 13, participants first review the evaluation criteria and then select their preferred image set according to four aspects: Prompt Fidelity, Identity Consistency, Style Consistency, and Image Quality. To ensure a consistent understanding of the evaluation criteria, we provide an illustrative example together with the rationale for the corresponding selection.

F Broader Impacts and Limitations

Broader Impacts. This work presents a training-free approach for consistent text-to-image generation, aiming to improve subject identity and appearance consistency across generated images. By avoiding task-specific fine-tuning, the proposed method reduces the computational cost associated with model adaptation and facilitates its use in applications such as personalized content creation, visual storytelling, and digital media generation.

Limitations. Although AdaShare improves multi-frame consistency without additional training, its adaptive feature sharing mechanism introduces additional inference overhead. Specifically, during diffusion sampling, AdaShare performs additional operations for subject-related information extraction, *ACOT Attention* optimization, and feature fusion to dynamically regulate information exchange across generated images. Consequently, these additional operations increase the computational cost of each sampling step relative to the underlying diffusion model, resulting in longer inference time. Future work will explore more efficient feature sharing strategies to reduce this computational overhead and improve inference efficiency.

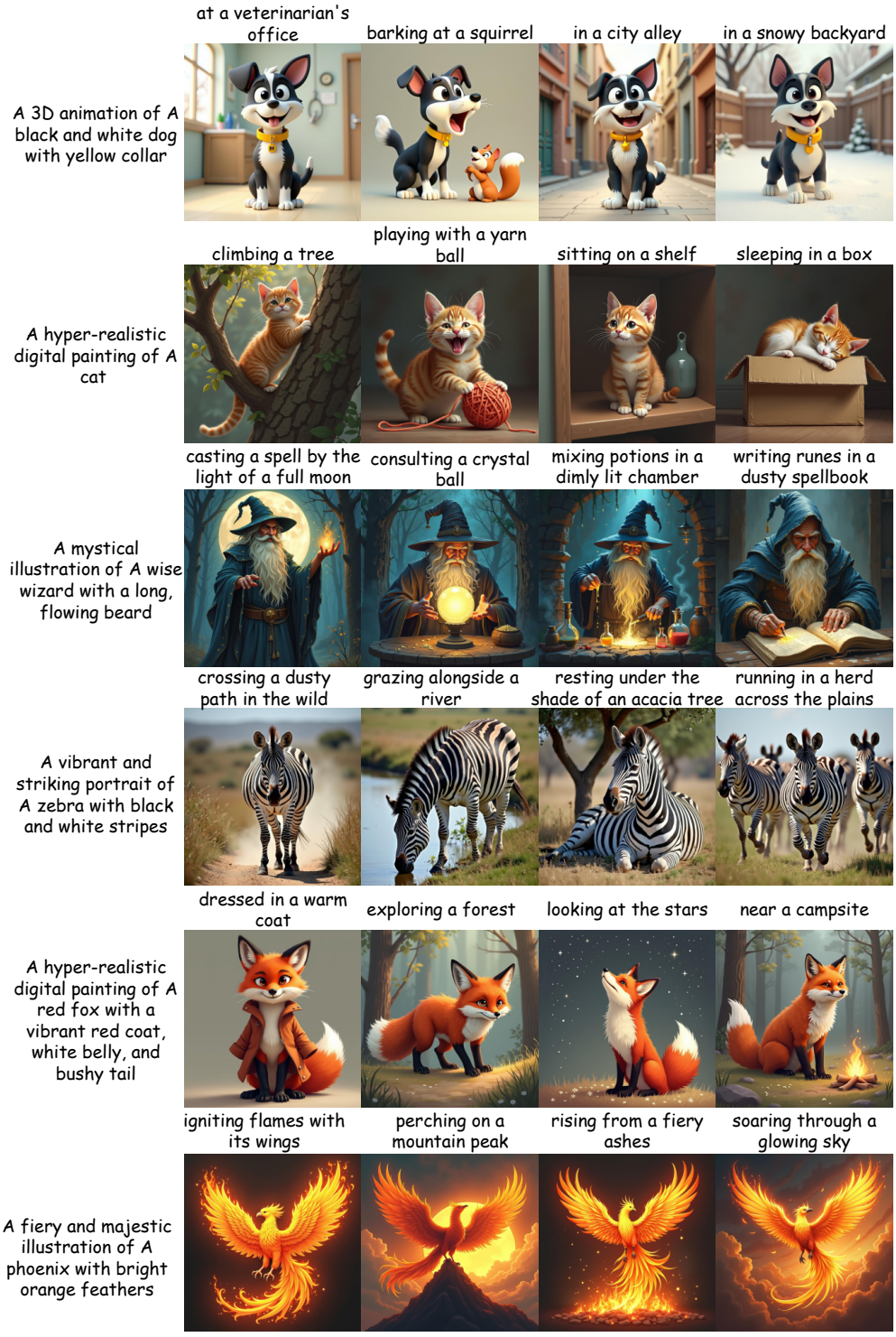


Figure 11: **Extension of AdaShare to FLUX.** FLUX, based on the DiT architecture, also maintains subject consistency and appearance consistency.

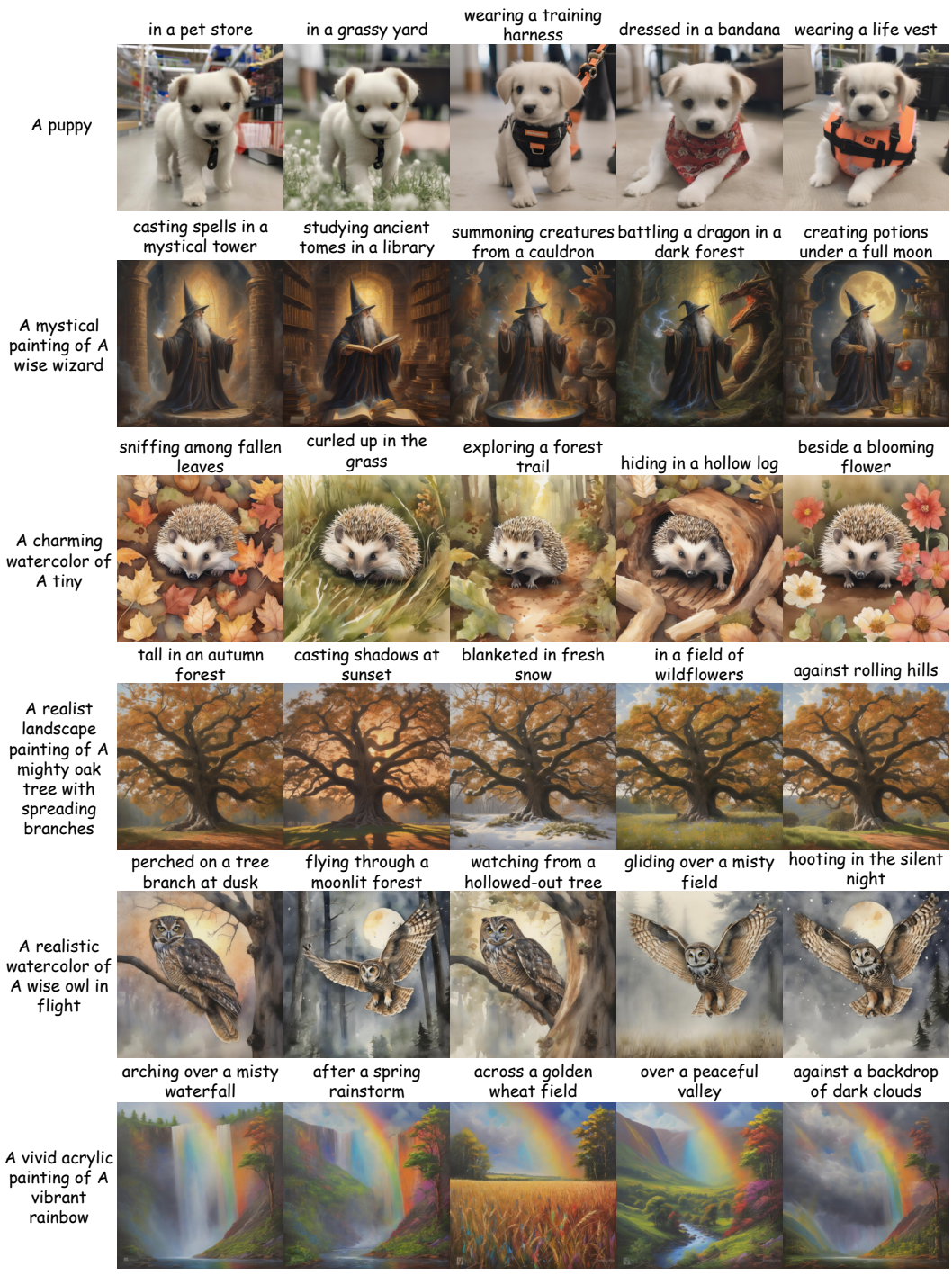
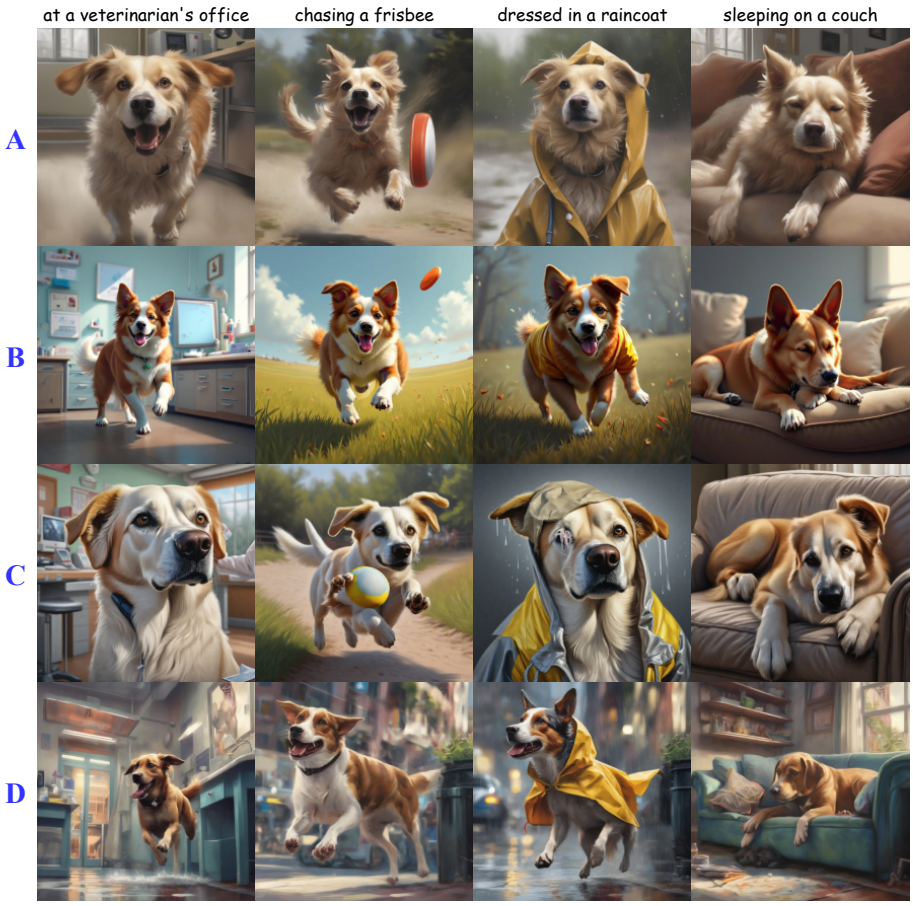


Figure 12: Additional qualitative results generated by AdaShare.

Evaluation Criteria

Prompt Fidelity: Determine the degree of match between each picture and its corresponding textual description.
Identity Consistency: Determine the extent to which the main subjects in the four pictures remain the same.
Style Consistency: Assess the extent to which the four images maintain a consistent visual style.
Image Quality: Assess the overall perceptual quality of each image, including visual realism, clarity, structural correctness, and the absence of noticeable artifacts, distortions, or unnatural details.

Example



Reasons for Selection

A: Maintains good performance in Prompt Fidelity, Identity Consistency, Style Consistency, and Image Quality.
B: Shows poor Image Quality and Identity Consistency.
C: Shows poor Image Quality.
D: Shows poor Identity Consistency.
Overall, A represents the best choice.

Figure 13: **User study questionnaire.** The Evaluation Criteria section presents four evaluation criteria: Prompt Fidelity, Identity Consistency, Style Consistency, and Image Quality. The Example section shows a sample interface for selection. The Reasons for Selection section provides the rationale for the selected option in this example.