MorphSIG: Subject-Driven Image Generation via Decoupled Anchoring and Feature Transport

Hailong Yan[1] **Yongrui Zhang**[1] **Xiangtao Zhang**[1] **Le Zhang**[2,1*]

[1] University of Electronic Science and Technology of China, Chengdu, China
[2] College of Computer Science and Software Engineering, Hohai University, Nanjing, China
yanhailong@std.uestc.edu.cn, lezhang@uestc.edu.cn



Figure 1: Qualitative comparison between MorphSIG and the baseline.

Abstract

We propose MorphSIG, a training-free framework that reformulates subject-driven image generation (SIG) as pseudo-video feature transport. MorphSIG contains two core modules: Progressive Decoupled Anchoring, which uses frequency-domain decoupling and structural scrambling during early denoising to break structural locking and enable a gradual transition from free composition to identity alignment; and Pseudo-flow Guided Feature Transport, which constructs a latent-space pseudo-flow field to mimic Image-to-Video temporal coherence and transport subject semantics precisely. To address stylistic homogeneity and the lack of ground truth in existing benchmarks, we further introduce a benchmark with diverse stylized subjects. Extensive experiments show that MorphSIG significantly improves pose diversity and subject fidelity over baselines without additional training. Overall, MorphSIG bridges static generation and dynamic propagation, providing a cost-effective paradigm for high-fidelity subject consistency. The Code and benchmark will be available here.

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

1 Introduction

Subject-driven image generation (SIG) aims to synthesize images that follow text prompts while preserving the details of reference subjects [27, 38, 16, 54, 12]. By enabling user-defined customization, SIG has broad applications in film production [58] and advertising [20].

In practical applications, SIG faces two major challenges: ❶ the "copy-paste pattern," where generated results inherit the reference pose and lack compositional freedom; and ❷ "subject inconsistency," where subject features degrade during diffusion, reducing identity stability.

As illustrated in Fig. 2, existing training-based SIG methods anchor identity through large-scale paired datasets [38, 15, 57, 54, 12] or the introduction of auxiliary encoding networks [59, 32, 49], yet these incur significant computational costs and are prone to overfitting. Training-free approaches [61, 31, 17, 7] typically inject subject features via attention-sharing mechanisms after inverting the reference image; however, they struggle to preserve identity due to the lack of consistency constraints. Another research [58, 52, 30, 63, 46, 10] attempts to leverage the dynamic priors of video models for image-level SIG, but suffers from low efficiency due to the generation of redundant transition frames.

To address these challenges, we analyze the underlying feature representations and inference paradigms. The "copy-paste pattern" arises from a pronounced "high-frequency shortcut effect" in Text&Image-to-Image (T&I2I) models, where high-frequency components from the reference image dominate early denoising and prematurely lock pose evolution. Meanwhile, the superior consistency of video models comes from temporally coherent feature transport, characterized by displacement mapping and semantic propagation across adjacent frames.

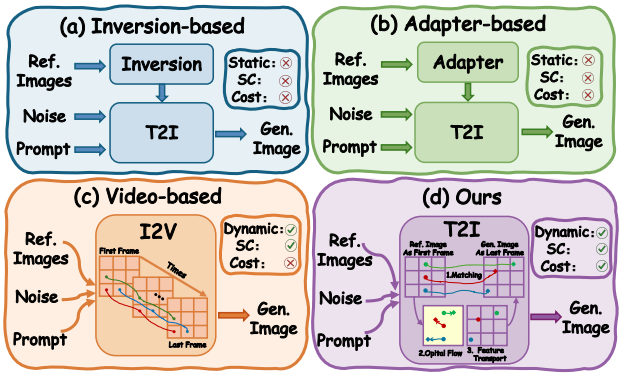


Figure 2: Comparison of different SIG paradigms. SC (Subject Consistency), and Cost (computational complexity).

To balance fidelity and diversity, we propose MorphSIG, a training-free SIG framework. First, to alleviate the "copy-paste" pattern, we introduce **Progressive Decoupled Anchoring**. By applying low-pass filtering and high-frequency structural permutation to reference Keys in joint self-attention, it removes global geometric bias while preserving essential feature statistics. Combined with time-dependent anchoring strength, this strategy unlocks compositional flexibility during early diffusion. Second, to improve consistency, we reformulate T&I2I as a "two-frame video generation" paradigm, treating the reference and target images as pseudo-first and pseudo-last frames. Based on this view, we propose **Pseudo-flow Guided Warped Feature Transport**, which simulates temporal coherence by warping reference features with pseudo-flow. Combining low-frequency Spherical Linear Interpolation and high-frequency Winner-Take-All enables effective transport of warped reference features, improving detail generation while preserving structural consistency.

We further introduce a benchmark with diverse styles and high-quality Ground Truth, enabling rigorous evaluation of fidelity and diversity in SIG. Our contributions are summarized as follows.

- We identify the "high-frequency shortcut" as the main cause of the "copy-paste" pattern in T&I2I, and redefine T&I2I as a "pseudo-video" task to simulate I2V temporal coherence without redundant computation.
- We propose Progressive Decoupled Anchoring to mitigate high-frequency shortcuts, and Pseudo-flow Guided Feature Transport to preserve subject consistency.
- We introduce a new SIG benchmark with diverse styles and high-quality Ground Truth, addressing the lack of objective metrics and stylistic diversity in existing datasets.
- As a training-free framework, MorphSIG achieves superior fidelity-diversity balance on both public and our proposed benchmarks.

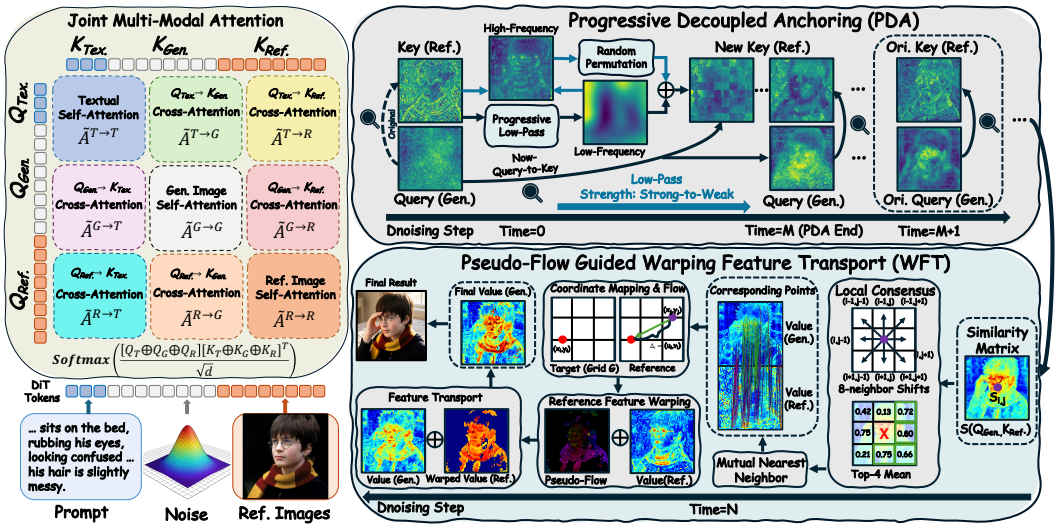


Figure 3: Overview of the proposed framework. MorphSIG consists of two modules: Progressive Decoupled Anchoring (PDA) and Warping Feature Transport (WFT). PDA suppresses structural shortcut copying by decoupling and progressively restoring reference anchors during denoising, enabling flexible composition. WFT estimates pseudo-flow to warp and transport reference features into the generation space, improving identity consistency under large pose and geometry variations.

2 Related Work

Consistent Text-to-Image Generation. While fine-tune approaches [38, 27, 22, 48, 43, 2, 64, 65] are capable of capturing visual details, they suffer from high computational overhead. To balance efficiency and performance, lightweight adapters [32, 62, 59] inject external guidance into frozen backbone; however, these often lack sufficient representation capacity. In contrast, training-free methods [16, 45, 31, 47, 28, 35, 36] leverage in-context learning for "plug-and-play" flexibility, but struggle with complex instructions, often failing to maintain textual consistency under large-scale pose variations.

Subject-driven Image Generation synthesizes reference images to precisely preserve subject identities [54, 12, 55]. For instance, DreamO [33] and USO [53] integrate reference images to generate outputs that mirror both input identity and aesthetics. While frameworks like OmniGen [56, 51] and others [25, 39, 13, 42] enhance identity decoupling, they often necessitate large-scale retraining, which limits their scalability. Conversely, training-free methods [61, 24, 11, 26] utilize optimized attention mechanisms to inject subject features across frames during inference.

Image Generation and Video. Recent video models [4, 5, 9, 23] have inspired using video priors for image tasks [37, 58, 63, 52]. Specifically, UniReal [10] and iMontage [19] treat image generation as video tasks to leverage world dynamics, while RealGeneral [30] unifies tasks via conditional frame prediction. To handle complexity, PLACID [44] uses spatio-temporal trajectories for feature constancy, and CoF-T2I [46] employs "Chain-of-Frames" for progressive refinement. Despite these gains, video-based models suffer from high overhead and latency. Conversely, we propose achieving high-quality generation by endowing static diffusion models with dynamic evolutionary capabilities.

3 Method

3.1 Preliminaries

Joint Multi-Modal Attention in T2I. Modern diffusion transformers like FLUX.2 utilizes a Joint Multi-Modal Attention mechanism that unifies text, latent noise, and reference images into a single 1D sequence. The model serializes the text prompt into X_{txt}, the target noise into X_{gen}, and N reference images into a concatenated stream $X_{\text{ref}} = [X_{\text{ref}}^{(1)}; \ldots; X_{\text{ref}}^{(N)}]$. To distinguish these "canvases," 4D coordinate IDs are assigned to every token, using discrete temporal offsets (T) for each reference

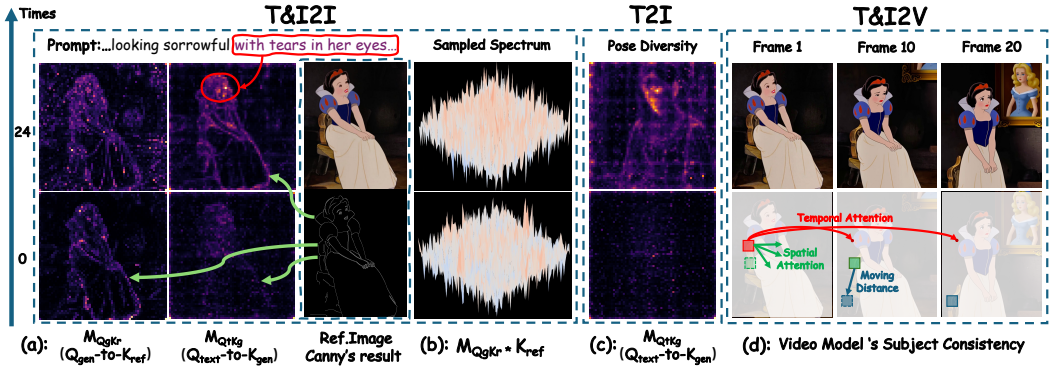


Figure 4: Analysis of High-frequency Shortcut in T&I2I and Temporal Feature Transport in I2V.

image to preserve instance integrity within the unified vector. In the core **Single-Stream Blocks**, these modalities are fused into a global sequence $X_{\text{all}} = [X_{\text{txt}}; X_{\text{gen}}; X_{\text{ref}}]$, where attention is computed as:

$$\text{Output} = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_h}}\right)V, \quad \{Q, K, V\} = \text{Linear}(X_{\text{all}}). \tag{1}$$

By treating references as tokens rather than external conditions, the joint attention matrix implicitly facilitates simultaneous self-attention, text-guidance, and multi-source reference injection.

3D Attention in I2V. Video DiT extend static DiTs by processing 4D video tensors $\mathcal{V} \in \mathbb{R}^{f\times c\times h\times w}$ to capture spatiotemporal dynamics. In each layer l, video tokens z_{video} and text tokens z_{text} are concatenated into an augmented sequence. The queries $\mathbf{Q}_{t,l}$ are formed by merging projections from all f frames and the text prompt: $\mathbf{Q}t, l = \text{Concat}(\mathbf{Q}t, l^1, \ldots, \mathbf{Q}t, l^f, \mathbf{Q}t, l^{\text{text}})$. The resulting global 3D attention matrix $\mathbf{A}_{t,l} = \text{Softmax}(\mathbf{Q}_{t,l}\mathbf{K}_{t,l}^\top/\sqrt{d})$ facilitates four key interactions: self-frame attention ($\mathbf{A}_{t,l}^{i,i}$) for 2D spatial correlations, cross-frame attention ($\mathbf{A}_{t,l}^{i,j}$) for inter-frame pixel relevance $\text{Attn}(I_i(p) \rightarrow I_j(q))$, and text-frame/self-text attention for semantic guidance. VDTs leverage inherent "world knowledge" to understand complex motion and temporal consistency [34].

3.2 Problems and Analysis

We analyze two key observations in T&I2I models, as shown in Fig.1: **Observation ❶: Copy-Paste Pattern.** Generated images rigidly inherit the reference pose. **Observation ❷: Subject Inconsistency.** Subject identity or features gradually degrade during generation. We first analyze and formally explain these issues, then propose targeted optimizations to balance fidelity and diversity.

Analysis for ❶: *The consistency gain in T&I2I comes at the expense of generation freedom.*

As shown in Fig.1(A), T&I2I models often directly replicate reference features, reducing diversity [21, 29, 57, 15]. We attribute this to a high-frequency shortcut effect: although diffusion models generally follow a coarse-to-fine generation process [60], T&I2I models capture high-frequency reference features too early to preserve identity consistency. This excessive coupling prematurely locks generation. Specifically, Fig.4(a) shows that Q_{gen} strongly activates edge contours in K_{ref} during early sampling. The spectrum in Fig.4(b) further shows that these high-frequency "anchors" constrain macro-scale pose evolution. As a result, the control ability of the text query Q_t is significantly weakened by the high-frequency shortcut, as illustrated in Fig.4(c).

Analysis for ❷: *The temporal consistency via first-frame feature transport.*

In I2V modeling, the attention for the latent sequence $\mathbf{z}_{video}$ explicitly constructs cross-frame attention by facilitating direct interaction between the query vectors $\mathbf{Q}_{t,l}^i$ of subsequent frames and the key vector $\mathbf{K}_{t,l}^1$ of the first frame, as illustrated in Fig.4(d). This architectural design establishes robust temporal correspondences, ensuring that visual priors from the initial frame are accurately propagated to subsequent frames, thereby maintaining identity stability while modeling temporal motion [41, 58]. Inspired by this, recent image-level methods [30, 10] adopt a "non-contiguous" frame generation architecture. By treating input and output images as a "pseudo-frame" sequence, they leverage the video model's inherent capacity for maintaining structural and feature consistency.

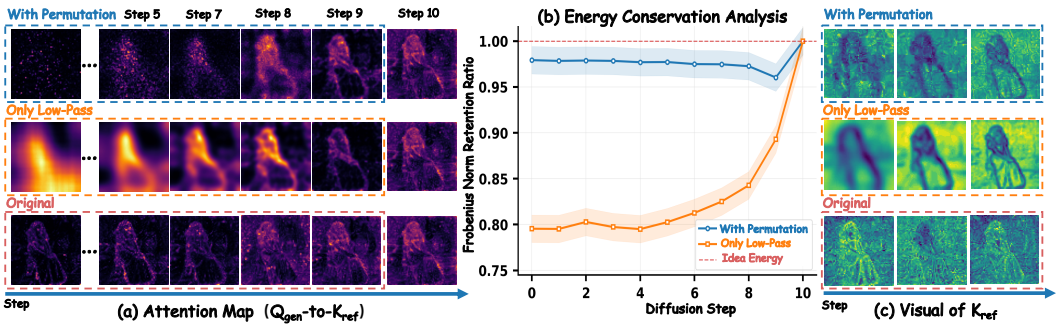


Figure 5: Synergistic Analysis of Spectral Decoupling and Structural Permutation.

3.3 Progressive Decoupled Anchoring (PDA)

Motivation. While T2I offer high compositional flexibility through noise synthesis, they struggle with subject consistency. Conversely, T&I2I anchor identity but often fall into "high-frequency shortcuts." Current solutions typically rely on computationally expensive paradigms, such as large-scale paired data augmentation [15, 57] or complex feedback optimization [12, 21]. We hope that an ideal T&I2I should balance the creative freedom of T2I with the robust consistency of T&I2I.

Spectral Decoupling (SD). In joint multi-modal attention, K_{ref} acts as a semantic anchor. Its high-frequency components often produce sharp attention peaks that lock generation to the reference geometry. To alleviate this pose-locking effect, we reshape the reference latent $\mathcal{M}_{ref} \in \mathbb{R}^{H\times W\times D}$ and apply frequency decoupling: $\mathcal{M}_{low} = \mathcal{G}(\mathcal{M}_{ref}, \sigma), \quad \mathcal{M}_{high} = \mathcal{M}_{ref} - \mathcal{M}_{low}$.

Where $\mathcal{G}$ denotes a Gaussian low-pass filter with a kernel scale σ. While $\mathcal{M}_{low}$ effectively circumvents high-frequency structural shortcuts, it also incurs a notable loss of feature energy. As illustrated in Fig. 5(b), an approximately 20% loss in *Frobenius Norm* significantly degrades the feature intensity of the pre-trained model, consequently compromising the fidelity of subject reconstruction.

Structural Permutation (SP). To compensate for information loss from filtering while eliminating geometric bias, we apply a block-level permutation $\mathcal{P}$ to the high-frequency residual $\mathcal{M}_{high}$:

$$\hat{\mathcal{M}}_{ref} = \mathcal{M}_{low} + \mathcal{P}(\mathcal{M}_{high}, p), \tag{2}$$

where p denotes the patch size. As visualized in Fig. 5(c), SP redistributes high-frequency features as "disassembled components" rather than merely erasing details. By disrupting global pose correlations while preserving original statistical properties and local details, SP avoids the identity blurring and feature collapse inherent in only low-pass schemes (which retain only 0.8 of the energy).

From the perspective of attention dynamics (Fig. 5(a)), the smooth distribution of only low-pass features (Fig. 5(c)) triggers attention diffusion, which impairs precise alignment. In contrast, while SP perturbs spatial coordinates, its block structures retain local gradient transitions. his high-contrast distribution (Fig. 5(a)) allows the model to bypass compositional constraints while accurately extracting critical details from the "feature pool," ensuring high-fidelity reconstruction.

Progressive Anchoring (PA). Leveraging the coarse-to-fine nature of diffusion denoising, we introduce a time-dependent coefficient $t \in (1 \to 0)$ to dynamically modulate anchoring intensity:

$$K_{ref} = \text{Flatten}\left(\mathcal{G}(\mathcal{M}_{ref}, \sigma(t)) + \mathcal{P}(\mathcal{M}_{ref} - \mathcal{G}(\mathcal{M}_{ref}, \sigma(t)), p)\right) \tag{3}$$

Macro-Evolution ($t \approx 1$): Low-pass filter suppresses high-frequency spatial anchors, forcing the model to synthesize global motion structures from the prompt Q_t within the "spectral blind zone."

Micro-Alignment ($t \to 0$): As the global manifold stabilizes, structural signals are reintroduced as local guides for identity reconstruction without disrupting the established pose.

3.4 Pseudo-Flow Guided Warping Feature Transport (WFT)

Motivation. The superior consistency of I2V stems from feature transport under temporal coherence [58, 6]. Existing methods [37, 63, 58] reformulate T2I as “video frame prediction” to inherit this property, but require generating redundant intermediate frames. We instead ask: **“Can T&I2I directly simulate the ‘I2V generation logic’ within T2I to enhance subject consistency?”**

Formulation: Consider a video sequence $(f_0, f_1, \ldots, f_n)$ as a continuous function in 2D space, where each frame f_j evolves from its predecessor f_{j-1} via a mapping T_j. The feature vector $\Phi_j(x)$ at coordinate x in the j-th frame is defined by the selective transport of prior features Φ_{j-1}:

$$\Phi_j(x) = \mathcal{R}\left(\Phi_{j-1}(T_j^{-1}(x)), \Delta\phi_j\right), \tag{4}$$

where T_j denotes the mapping function, and $\mathcal{R}$ represents the transport operator. It propagates semantic information from the reference frame to the current frame according to the spatial mapping, with $\Delta\phi_j$ accounting for frame-specific residual updates. This formulation reveals that temporal coherence fundamentally arises from the combination of spatial feature transport and local updates.

In I2V, The 3D attention implicitly realizes the above process through Query-Key matching [34]: $A(x, y) \propto \langle Q_j(x), K_{j-1}(y)\rangle$ which can be interpreted as a soft correspondence distribution enabling cross-frame non-rigid alignment. Functionally, it resembles a probabilistic flow field [14, 41].

Motivated by this observation, we reinterpret T&I2I architectures (e.g., FLUX.2) as a degenerate two-frame I2V process, where the reference and target images correspond to pseudo-first frame f_0 and pseudo-final frame f_1. We then construct a pseudo-flow mapping T within T&I2I to model semantic correspondence: $\Phi_1(x) = \mathcal{R}(\Phi_0(T^{-1}(x)), \Delta\phi(x))$, where pseudo-flow performs semantic transport and the residual term injects text-conditioned novel semantics.

Method. Unlike methods that rely on external models to establish correspondences [3], we compute the global pairwise similarity $S \in \mathbb{R}^{L_{gen}\times L_{ref}}$ between object Q_{gen} and reference K_{ref}:

$$S_{i,j} = \frac{\langle Q_{gen,i}, K_{ref,j}\rangle}{\|Q_{gen,i}\|\|K_{ref,j}\|}. \tag{5}$$

To suppress redundant matches and enforce spatial coherence, we introduce a two-stage constraint:

❶ **Local Consensus (LC).** To suppress semantic mismatches caused by repetitive textures, we introduce a neighborhood continuity assumption, i.e., correctly matched points should exhibit consistent local motion patterns. Specifically, we apply 8-neighbor shifts to $S_{i,j}$ and compute the mean of the Top-4 neighborhood responses as the local weight W_{local} :

$$W_{local} = \text{Top-K}_{mean}\left(\{S(x+\Delta x, y+\Delta y) \mid (\Delta x, \Delta y) \in \text{Shifts}\}, k=4\right) \tag{6}$$

Using Top-4 restricts responses to valid regions near boundaries and occlusions. The refined matrix $\hat{S} = S \odot W_{local}$ effectively suppresses isolated mismatches.

❷ **Mutual Nearest Neighbor (MNN).** To further improve matching robustness, we impose a bidirectional Top-K constraint. A correspondence is considered valid only when target position i and reference position j are mutual Top-K matches: $M_{valid} = M_{g\to r}(K) \cap M_{r\to g}(K)$. The optimal correspondence is then selected as $j^* = \arg\max_j(\hat{S}_{i,j} \cdot M_{valid})$, which is mapped back to 2D coordinates. The offset relative to the sampling grid $\mathcal{G}$ yields the pseudo-flow field: $\mathcal{F} \in \mathbb{R}^{2\times H\times W}$.

Reference Feature Warping. Given $\mathcal{F}$, the reference value V_{ref} are aligned to the object feature space. Specifically, the sampling coordinates are defined as $\mathcal{G}_{sample} = \mathcal{G} + \mathcal{F}$. and bilinear resampling is performed in the normalized coordinate space $[-1, 1]$, with out-of-bound regions zero-padded to avoid boundary stretching artifacts. Meanwhile, the final constraint mask is constructed from M_{valid} and $\hat{S}$ to filter mismatched and occluded regions: $V_{warp} = \text{Sample}(V_{ref}, \mathcal{G}_{sample}) \odot M_{final}$.

Warped Feature Transport. Directly injecting warped reference features V_{warp} into V_{gen} causes distribution mismatch. Although V_{warp} preserves reference semantics, non-rigid deformation and interpolation introduce deviations from the V_{gen} in local details and feature statistics. Thus, we formulate the transport process as $\mathcal{R}(V_{gen}, V_{warp})$, where $\mathcal{R}$ is a spatial frequency-decoupled operator that separately models structural consistency and detail fidelity.

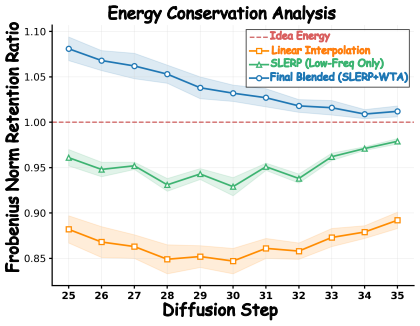


Figure 6: Feature Energy Analysis.

❶ **Low-Frequency SLERP.** Low-frequency components encode global structures such as contours and color distribution. We transport V_{warp}^{low} to V_{gen}^{low} using spherical linear interpolation (SLERP) [1]. ❷ **High-Frequency Winner-Takes-All.** High-frequency contain texture

Table 1: Quantitative Results of Various SIG Methods on ViStoryBench [66]. CI: Character Interaction, IA: Individual Action; The first, second, and third values are highlighted.

Methods	Venue	CSD↑		CIDS↑		PA↑					AES↑	CP↓	INC↑
		Cross	Self	Cross	Self	Scene	Shot	CI	IA	Avg.			
DreamO [33]	SIGGA'25	0.489	0.584	0.542	0.643	2.99	2.21	1.72	1.65	2.14	5.31	0.533	9.21
DiptychP [40]	CVPR'25	0.394	0.462	0.417	0.536	2.73	2.04	1.61	1.50	1.97	4.60	0.524	13.18
XVerse [8]	NIPS'25	0.229	0.461	0.342	0.531	2.96	2.37	1.89	1.73	2.23	4.92	0.519	13.67
UNO [54]	ICCV'25	0.513	0.576	0.528	0.632	3.08	2.52	2.07	1.91	2.39	4.94	0.542	8.27
OmniGen2 [51]	arXiv'25	0.446	0.546	0.460	0.587	3.12	2.46	1.80	1.64	2.27	5.19	0.531	9.87
Scone [50]	CVPR'26	0.511	0.539	0.501	0.585	3.11	2.40	2.18	1.69	2.35	5.22	0.517	10.60
UMO [12]	CVPR'26	0.514	0.635	0.555	0.648	3.25	2.65	2.13	1.93	2.49	5.05	0.536	10.35
MOSAIC [39]	ICLR'26	0.268	0.463	0.377	0.531	3.14	2.43	2.31	1.96	2.46	5.36	0.548	13.56
FLUX.2-4B	-	0.489	0.531	0.576	0.637	3.27	2.54	2.27	1.99	2.52	5.42	0.537	10.97
Ours(FLUX.2-4B)	-	**0.521**	**0.598**	**0.605**	**0.653**	**3.32**	**2.58**	**2.35**	**2.08**	**2.58**	**5.39**	**0.516**	**13.88**

details and are sensitive to alignment errors. Direct transport may introduce ghosting or blur artifacts. Therefore, we adopt a Winner-Takes-All (WTA) strategy. The final generated value is computed as:$V_{gen} = \text{SLERP}(V_{gen}^{low}, V_{warp}^{low}) + \text{argmax}_{V \in \{V_{warp}^{high}, V_{gen}^{high}\}} \|V\|$.

❸**Discuss.** As shown in Fig. 6, we analyze energy evolution using the Frobenius-norm retention ratio. Compared with linear interpolation, SLERP substantially mitigates energy attenuation, thereby better preserving the low-frequency information established in early denoising stages. Moreover, the synergy of SLERP and WTA yields a final feature energy slightly above the ideal expectation (ratio > 1.0), providing additional support for reconstructing high-frequency details in the mid-to-late denoising stages.

4 Experiments

Implementation Details. We implement our method on FLUX.2-Klein-4B, a mainstream T&I2I framework with a DiT-based joint self-attention architecture. Progressive spectral decoupling is applied in the first 20% of denoising steps, while pseudo-flow-guided feature transport is injected during the intermediate phase, i.e., 50%–70% of steps, when semantic structures are stabilized but fine-grained identity details are still refined. All experiments are conducted at 1024×1024 resolution on NVIDIA A100 40GB GPUs. Unless otherwise specified, hyperparameters are fixed across benchmarks and prompts. Detailed hyperparameters are provided in the Appendix.

Baselines. We compare our method with recent SOTA SIG approaches covering both training-based and training-free paradigms: Scone [50], UMO [12], UNO [54], MOSAIC [39], OmniGen2 [51], DreamO [33], DiptychP [40], and XVerse [8]. These methods improve subject consistency through large-scale paired training, preference optimization, explicit identity supervision, or inference-time attention/latent feature manipulation without finetuning. For fair comparison, all baselines use official checkpoints and default hyperparameters.

Evaluation Metrics. We evaluate our framework on the public VistoryBench [66] and our proposed SIGBench. VistoryBench includes 1,317 cases and 344 characters. VistoryBench measures subject consistency using Character Identification Similarity (CIDS) and Character Style Similarity (CSD), prompt alignment using a multimodal large language model, aesthetic quality (AES) using Aesthetic Predictor V2.5, and diversity using Inception Score (INC) and Copy-Paste (CP) Detection metrics. On SIGBench, we use CLIP-I and DreamSim [18] for subject consistency, the proposed Comprehensive Diversity Score (CDS) for diversity, CLIP-T and PA for prompt alignment, and Aesthetic Predictor V2.5 for aesthetic quality. More details are provided in the Appendix.

Quantitative Results. Table 1 shows that our method achieves the excellent overall performance on VistoryBench. It ranks first on CSD-Cross, CIDS-Cross, CIDS-Self, and average PA, demonstrating stronger subject consistency and prompt alignment. Compared with baseline, it improves CIDS-Cross from 0.576 to 0.605 and average PA from 2.52 to 2.58, indicating that WFT improves identity-preserving feature reuse while PDA reduces reference-induced structural shortcuts. It also achieves the lowest Copy-Paste score of 0.516 and the highest Inception Score of 13.88, showing greater diversity and less reference copying.

Table 2 reports results on SIGBench, where our method ranks first on DreamSim-GT, PA-Avg., CLIP-I, and CDS. Compared with FLUX.2-4B, it improves PA-Avg. from 3.06 to 3.22, CLIP-I from 0.847

Table 2: Quantitative Results of Various SIG Methods on SIGBench. VRAM: Using GPU memory during inference. The first, second, and third values are highlighted.

Methods	Venue	DreamSim↓		PA↑				CLIP-T↑	CLIP-I↑	CDS↑	AES↑	VRAM↓
		GT	Ref	Scene	CI	IA	Avg.					
DreamO [33]	SIGGA'25	0.273	0.596	3.04	2.43	2.07	2.51	0.361	0.834	0.373	6.00	9.09
DiptychP [40]	CVPR'25	0.459	0.651	2.72	2.16	2.45	2.44	0.332	0.696	0.339	5.24	30.01
XVerse [8]	NIPS'25	0.339	0.660	2.95	2.53	2.12	2.53	0.358	0.767	0.366	6.09	37.39
UNO [54]	ICCV'25	0.315	0.604	3.21	2.59	2.28	2.69	0.345	0.802	0.366	5.64	35.09
OmniGen2 [51]	arXiv'25	0.287	0.607	3.28	2.67	2.41	2.78	0.363	0.822	0.380	5.47	9.92
Scone [50]	CVPR'26	0.261	0.601	3.51	2.88	2.57	2.98	0.362	0.838	0.383	5.83	30.43
UMO [12]	CVPR'26	0.296	0.602	3.38	3.01	2.44	2.94	0.348	0.815	0.375	5.97	33.78
MOSAIC [39]	ICLR'26	0.281	0.621	3.43	2.70	2.56	2.89	0.354	0.824	0.401	5.82	35.79
FLUX.2-4B	-	0.225	0.643	3.40	3.15	2.63	3.06	0.354	0.847	0.412	6.33	12.21
Ours (FLUX.2-4B)	-	**0.221**	**0.625**	**3.47**	**3.34**	**2.86**	**3.22**	**0.362**	**0.854**	**0.429**	**6.32**	**13.44**

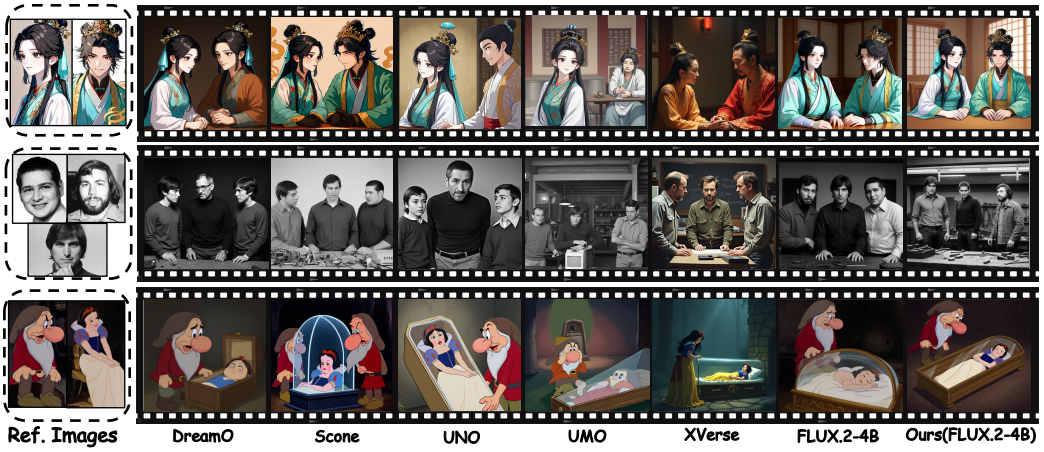


Figure 7: Qualitative Comparison on ViStoryBench Across Diverse Characters and Styles.

to 0.854, and CDS from 0.412 to 0.429, showing better prompt alignment, subject consistency, and diversity. These results support our design: PDA reduces pose copying and enhances compositional flexibility, while WFT transports identity features via pseudo-flow alignment. With only minor VRAM overhead over FLUX.2-4B, our method remains more efficient than most training-based baselines and achieves superior overall performance.

Qualitative Results. Fig. 7 and 8 show qualitative comparisons on ViStoryBench. Existing methods often exhibit identity drift, missing subjects, style inconsistency, or direct copying of the reference pose. In contrast, our method better preserves character identity and visual style while allowing prompt-driven changes in pose, interaction, and scene composition.

In multi-character cases, our method maintains more stable identities for different subjects and produces more coherent interactions. These results indicate that PDA reduces reference-structure shortcuts, while WFT restores identity-related details through feature transport.

Fig. 9 shows qualitative comparisons on SIGBench, where each case provides reference subjects and a ground-truth target image. Compared with existing methods, our method better preserves subject identity and fine-grained attributes while matching the target composition more closely. Baselines often miss objects, alter subject appearance, or over-copy the reference layout, especially in multi-subject and large-pose-change scenarios. Our results show more accurate subject-object relationships, more stable appearance preservation, and better alignment with the GT scene structure.

Table 3: Human preference comparison among different methods.

Method	**DreamO**	**DiptychP**	**XVerse**	**UNO**	**OmniGen2**	**Scone**	**UMO**	**MOSAIC**	**FLUX.2-4B**	**Ours**
Percent (%)↑	2.79	1.41	0.62	10.19	3.03	12.84	15.95	4.48	22.75	**26.31**

User Study. To complement automatic metrics, we conduct a human evaluation on 50 randomly sampled cases from the two benchmarks. For each case, all competing methods generate corresponding image sequences. We invite 20 participants, each evaluating 20 randomly selected results with the reference images and prompts. They are asked to select the preferred sequence based on subject



Figure 8: Qualitative Comparison on Multi-Character Story Generation in ViStoryBench.



Figure 9: Qualitative Comparison on SIGBench with Ground Truth.

consistency, pose diversity, and overall image quality. As shown in Table 3, our method achieves the highest preference rate. The results further demonstrate that MorphSIG produces more consistent and visually appealing generations while maintaining stronger structural diversity. Details of the protocol and interface are provided in the Appendix.

Ablation Study. Tables 4 and 5 report the ablation results of PDA and WFT. Removing PDA degrades CIDS, INC, and PA, indicating excessive reliance on reference structures and reduced flexibility in text-driven pose and layout generation. Specifically, removing SP weakens identity-preserving details, while removing SD reduces pose diversity, demonstrating their respective roles in identity retention and structural decoupling. Removing WFT causes a significant drop in CIDS, highlighting the importance of pseudo-flow-guided feature transport for maintaining subject consistency under large structural variations. Further removing MNN, LF-SLERP, or HF-WTA consistently lowers performance, showing that bidirectional matching, low-frequency fusion, and high-frequency detail selection jointly improve structural stability and texture fidelity.

Table 4: The ablation study of PDA.

	CIDS↑	INC↑	PA↑	AES↑
w/o PDA	0.614	13.41	2.37	5.30
w/o SP	0.556	13.64	**2.72**	5.33
w/o SD	0.618	13.57	2.39	5.36
High-pass only	0.485	12.81	1.92	5.05
Full PDA	**0.629**	**13.88**	2.58	**5.39**

Fig. 10 illustrates the visual differences under various ablation settings. Without PDA, the generated results tend to replicate the original pose, limiting the controllability of text-driven generation. Removing SP or SD further causes identity-detail degradation and insufficient structural disentanglement, respectively. For WFT-related ablations, removing WFT or MNN introduces local misalignment and identity drift, while removing LF-SLERP or HF-WTA results in unstable structures or blurred textures. In contrast, the full model better preserves subject consistency while generating novel poses and scenes aligned with the

Table 5: The ablation study of WFT.

	CIDS↑	INC↑	PA↑	AES↑
w/o WFT	0.541	13.47	**2.64**	5.25
w/o MNN	0.578	13.32	2.47	5.18
w/o HF-WTA	0.601	13.65	2.52	5.14
w/o LF-SLERP	0.584	13.30	2.34	5.19
ADD	0.561	12.93	2.16	5.05
Full WFT	**0.629**	**13.88**	2.58	**5.39**

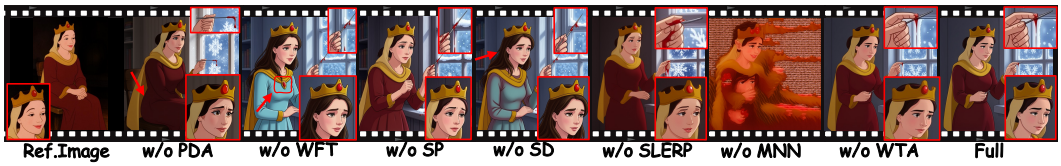


Figure 10: Qualitative Comparison of Ablation Study.

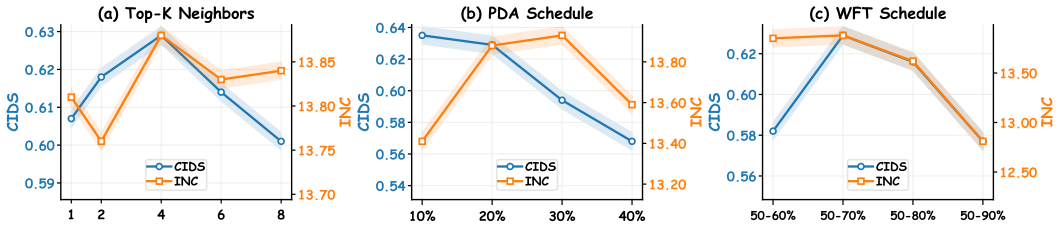


Figure 11: **Ablation analysis.** (a) Effect of neighborhood size K in Local Consensus. (b, c) Sensitivity of PDA and WFT scheduling ranges.

text prompt, demonstrating the complementary effects of PDA and WFT in harmonizing identity retention with generative diversity.

Local Consensus Analysis. As shown in Fig. 11(a), selecting $K = 4$ in Local Consensus achieves the best balance between identity fidelity (CIDS) and diversity (INC). A smaller K admits spurious matches from texture ambiguity, whereas a larger K over-suppresses valid correspondences by incorporating noisy boundary neighbors.

Scheduling Sensitivity Analysis. As shown in Fig. 11(b, c), our default schedule achieves the optimal balance between identity fidelity and generation diversity. Terminating PDA prematurely (10%) locks macro-composition via high-frequency shortcuts, whereas prolonging it (>20%) impairs fine-grained detail alignment. Similarly, WFT requires sufficient propagation without over-intervention: truncating it early (<70%) hinders detail consistency, while extending it into late denoising (>80%) introduces distribution mismatch and artifacts. Thus, our setting (PDA: 20%, WFT: 50%–70%) reliably optimizes performance.

5 Conclusion

We presented MorphSIG, a training-free framework for subject-driven image generation. By identifying high-frequency reference shortcuts as a major cause of copy-paste artifacts, we proposed Progressive Decoupled Anchoring to relax structural locking while preserving identity-related statistics. We further reformulated T&I2I as pseudo-video generation and introduced Pseudo-flow Guided Feature Transport to improve subject consistency through latent-space feature transport without generating intermediate frames. In addition, we built a new SIG benchmark with diverse styles and high-quality ground truth. Experiments on public and proposed benchmarks show that MorphSIG achieves a strong balance between identity fidelity and generation diversity. These results demonstrate the effectiveness of frequency-decoupled conditioning and pseudo-flow-guided transport for controllable and consistent subject-driven generation.

6 Acknowledgements and Disclosures

This research is supported by the National Key R&D Program of China (No. 2024YFE0100700). The authors declare that they have no competing interests.

References

[1] Amirhossein Alimohammadi, Aryan Mikaeili, Sauradip Nag, Negar Hassanpour, Andrea Tagliasacchi, and Ali Mahdavi-Amiri. Cora: Correspondence-aware image editing using few step diffusion. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–11, 2025.

[2] Guangrui Bai, Hailong Yan, Wenhai Liu, Yahui Deng, and Erbao Dong. Towards lightest low-light image enhancement architecture for mobile devices. *Expert Systems with Applications*, 296:129125, 2026.

[3] Qingyan Bai, Hao Ouyang, Yinghao Xu, Qiuyu Wang, Ceyuan Yang, Ka Leong Cheng, Yujun Shen, and Qifeng Chen. Edicho: Consistent image editing in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15277–15287, 2025.

[4] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

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

[6] Ryan Burgert, Yuancheng Xu, Wenqi Xian, Oliver Pilarski, Pascal Clausen, Mingming He, Li Ma, Yitong Deng, Lingxiao Li, Mohsen Mousavi, et al. Go-with-the-flow: Motion-controllable video diffusion models using real-time warped noise. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13–23, 2025.

[7] Kelvin CK Chan, Yang Zhao, Xuhui Jia, Ming-Hsuan Yang, and Huisheng Wang. Improving subject-driven image synthesis with subject-agnostic guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6733–6742, 2024.

[8] Bowen Chen, Mengyi Zhao, Haomiao Sun, Li Chen, Xu Wang, Kang Du, and Xinglong Wu. Xverse: Consistent multi-subject control of identity and semantic attributes via dit modulation. *arXiv preprint arXiv:2506.21416*, 2025.

[9] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023.

[10] Xi Chen, Zhifei Zhang, He Zhang, Yuqian Zhou, Soo Ye Kim, Qing Liu, Yijun Li, Jianming Zhang, Nanxuan Zhao, Yilin Wang, et al. Unireal: Universal image generation and editing via learning real-world dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12501–12511, 2025.

[11] Yuzhuo Chen, Zehua Ma, Jianhua Wang, Kai Kang, Shunyu Yao, and Weiming Zhang. Lamic: Layout-aware multi-image composition via scalability of multimodal diffusion transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 3174–3182, 2026.

[12] Yufeng Cheng, Wenxu Wu, Shaojin Wu, Mengqi Huang, Fei Ding, and Qian He. Umo: Scaling multi-identity consistency for image customization via matching reward. *arXiv preprint arXiv:2509.06818*, 2025.

[13] Yusuf Dalva, Hidir Yesiltepe, and Pinar Yanardag. Lorashop: Training-free multi-concept image generation and editing with rectified flow transformers. *arXiv preprint arXiv:2505.23758*, 2025.

[14] Giannis Daras, Weili Nie, Karsten Kreis, Alexandros G Dimakis, Morteza Mardani, Nikola B Kovachki, and Arash Vahdat. Warped diffusion: Solving video inverse problems with image diffusion models. *Advances in Neural Information Processing Systems*, 37:101116–101143, 2024.

[15] Yufan Deng, Yuanyang Yin, Xun Guo, Yizhi Wang, Jacob Zhiyuan Fang, Shenghai Yuan, Yiding Yang, Angtian Wang, Bo Liu, Haibin Huang, et al. Magref: Masked guidance for any-reference video generation with subject disentanglement. *arXiv preprint arXiv:2505.23742*, 2025.

[16] Ganggui Ding, Canyu Zhao, Wen Wang, Zhen Yang, Zide Liu, Hao Chen, and Chunhua Shen. Freecustom: Tuning-free customized image generation for multi-concept composition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9089–9098, 2024.

[17] Haoran Feng, Zehuan Huang, Lin Li, and Lu Sheng. Personalize anything for free with diffusion transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 3921–3929, 2026.

[18] 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. *arXiv preprint arXiv:2306.09344*, 2023.

[19] Zhoujie Fu, Xianfang Zeng, Jinghong Lan, Xinyao Liao, Cheng Chen, Junyi Chen, Jiacheng Wei, Wei Cheng, Shiyu Liu, Yunuo Chen, et al. imontage: Unified, versatile, highly dynamic many-to-many image generation. *arXiv preprint arXiv:2511.20635*, 2025.

[20] Yifan Gao, Zihang Lin, Chuanbin Liu, Min Zhou, Tiezheng Ge, Bo Zheng, and Hongtao Xie. Postermaker: Towards high-quality product poster generation with accurate text rendering. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8083–8093, 2025.

[21] Anujraaj Argo Goyal, Guocheng Gordon Qian, Huseyin Coskun, Aarush Gupta, Himmy Tam, Daniil Ostashev, Ju Hu, Dhritiman Sagar, Sergey Tulyakov, Kfir Aberman, et al. Preventing shortcuts in adapter training via providing the shortcuts. *arXiv preprint arXiv:2510.20887*, 2025.

[22] Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Yunpeng Chen, Zihan Fan, Wuyou Xiao, Rui Zhao, Shuning Chang, Weijia Wu, et al. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. *Advances in Neural Information Processing Systems*, 36:15890–15902, 2023.

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

[24] Jinghao Hu, Yuhe Zhang, GuoHua Geng, Kang Li, and Han Zhang. Storytailor: A zero-shot pipeline for action-rich multi-subject visual narratives. *arXiv preprint arXiv:2602.21273*, 2026.

[25] Xirui Hu, Jiahao Wang, Hao Chen, Weizhan Zhang, Benqi Wang, Yikun Li, and Haishun Nan. Dynamicid: Zero-shot multi-id image personalization with flexible facial editability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10549–10559, 2025.

[26] Hao Kang, Stathi Fotiadis, Liming Jiang, Qing Yan, Yumin Jia, Zichuan Liu, Min Jin Chong, and Xin Lu. Flux already knows–activating subject-driven image generation without training. *arXiv preprint arXiv:2504.11478*, 2025.

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

[28] Mingxiao Li, Mang Ning, and Marie-Francine Moens. Consistent story generation: Unlocking the potential of zigzag sampling. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[29] Shuang Li, Chao Deng, Hang Chen, Liqun Liu, Zhenyu Hu, Te Cao, Mengge Xue, Yuan Chen, Peng Shu, Huan Yu, et al. Disentangling to re-couple: Resolving the similarity-controllability paradox in subject-driven text-to-image generation. *arXiv preprint arXiv:2604.00849*, 2026.

[30] Yijing Lin, Mengqi Huang, Shuhan Zhuang, and Zhendong Mao. Realgeneral: Unifying visual generation via temporal in-context learning with video models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14994–15004, 2025.

[31] Yohai Mazuz, Janna Bruner, and Lior Wolf. Consistyle: Style diversity in training-free consistent t2i generation. *ACM Transactions on Graphics (TOG)*, 44(6):1–16, 2025.

[32] 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, pages 4296–4304, 2024.

[33] Chong Mou, Yanze Wu, Wenxu Wu, Zinan Guo, Pengze Zhang, Yufeng Cheng, Yiming Luo, Fei Ding, Shiwen Zhang, Xinghui Li, et al. Dreamo: A unified framework for image customization. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pages 1–12, 2025.

[34] Jisu Nam, Soowon Son, Dahyun Chung, Jiyoung Kim, Siyoon Jin, Junhwa Hur, and Seungryong Kim. Emergent temporal correspondences from video diffusion transformers. *arXiv preprint arXiv:2506.17220*, 2025.

[35] Xinyu Pu, Hongsong Wang, Jie Gui, and Pan Zhou. Drag as evidence: Motion-grounded latent recomposition for drag-based editing. *arXiv preprint arXiv:2609.36755*, 2026.

[36] Xinyu Pu, Hongsong Wang, Jie Gui, and Pan Zhou. Dragging with geometry: From pixels to geometry-guided image editing. In *International Conference on Learning Representations*, volume 2026, pages 151958–151984, 2026.

[37] Noam Rotstein, Gal Yona, Daniel Silver, Roy Velich, David Bensaïd, and Ron Kimmel. Pathways on the image manifold: Image editing via video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7857–7866, 2025.

[38] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023.

[39] Dong She, Siming Fu, Mushui Liu, Qiaoqiao Jin, Hualiang Wang, Mu Liu, and Jidong Jiang. Mosaic: Multi-subject personalized generation via correspondence-aware alignment and disentanglement. *arXiv preprint arXiv:2509.01977*, 2025.

[40] Chaehun Shin, Jooyoung Choi, Heeseung Kim, and Sungroh Yoon. Large-scale text-to-image model with inpainting is a zero-shot subject-driven image generator. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7986–7996, 2025.

[41] Yeji Song, Wonsik Shin, Junsoo Lee, Jeesoo Kim, and Nojun Kwak. Save: Protagonist diversification with s tructure a gnostic v ideo e diting. In *European Conference on Computer Vision*, pages 41–57. Springer, 2024.

[42] Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14940–14950, 2025.

[43] Bowen Tang, Tao Wang, Miao Zhang, Xin Yu, Jinwei Chen, Bo Li, and Kaihao Zhang. Evolutionary multi-agent collaboration for real-world video face restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8175–8184, 2026.

[44] Gemma Canet Tarrés, Manel Baradad, Francesc Moreno-Noguer, and Yumeng Li. Placid: Identity-preserving multi-object compositing via video diffusion with synthetic trajectories. *arXiv preprint arXiv:2602.00267*, 2026.

[45] 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.

[46] Chengzhuo Tong, Mingkun Chang, Shenglong Zhang, Yuran Wang, Cheng Liang, Zhizheng Zhao, Ruichuan An, Bohan Zeng, Yang Shi, Yifan Dai, et al. Cof-t2i: Video models as pure visual reasoners for text-to-image generation. *arXiv preprint arXiv:2601.10061*, 2026.

[47] Mengyu Wang, Henghui Ding, Jianing Peng, Yao Zhao, Yunpeng Chen, and Yunchao Wei. Characonsist: Fine-grained consistent character generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16058–16067, 2025.

[48] Tao Wang, Peiwen Xia, Bowen Tang, Jinwei Chen, Kaihao Zhang, and Bi Li. Fair and faithful: A diffusion-enhanced dataset and hybrid state-space mamba for face super-resolution. In *European Conference on Computer Vision*, pages 281–300, 2026.

[49] Xierui Wang, Siming Fu, Qihan Huang, Wanggui He, and Hao Jiang. Ms-diffusion: Multi-subject zero-shot image personalization with layout guidance. *arXiv preprint arXiv:2406.07209*, 2024.

[50] Yuran Wang, Bohan Zeng, Chengzhuo Tong, Wenxuan Liu, Yang Shi, Xiaochen Ma, Hao Liang, Yuanxing Zhang, and Wentao Zhang. Scone: Bridging composition and distinction in subject-driven image generation via unified understanding-generation modeling. *arXiv preprint arXiv:2512.12675*, 2025.

[51] Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025.

[52] Jay Zhangjie Wu, Xuanchi Ren, Tianchang Shen, Tianshi Cao, Kai He, Yifan Lu, Ruiyuan Gao, Enze Xie, Shiyi Lan, Jose M Alvarez, et al. Chronoedit: Towards temporal reasoning for image editing and world simulation. *arXiv preprint arXiv:2510.04290*, 2025.

[53] Shaojin Wu, Mengqi Huang, Yufeng Cheng, Wenxu Wu, Jiahe Tian, Yiming Luo, Fei Ding, and Qian He. Uso: Unified style and subject-driven generation via disentangled and reward learning. *arXiv preprint arXiv:2508.18966*, 2025.

[54] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18682–18692, 2025.

[55] Tao Wu, Yibo Jiang, Yehao Lu, Zhizhong Wang, Zeyi Huang, Zequn Qin, and Xi Li. Multicrafter: High-fidelity multi-subject generation via disentangled attention and identity-aware preference alignment. *arXiv preprint arXiv:2509.21953*, 2025.

[56] Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13294–13304, 2025.

[57] Hengyuan Xu, Wei Cheng, Peng Xing, Yixiao Fang, Shuhan Wu, Rui Wang, Xianfang Zeng, Daxin Jiang, Gang Yu, Xingjun Ma, et al. Withanyone: Towards controllable and id consistent image generation. *arXiv preprint arXiv:2510.14975*, 2025.

[58] Hailong Yan, Shice Liu, Tao Wang, Xiangtao Zhang, Yijie Zhong, Jinwei Chen, Le Zhang, and Bo Li. Animeagent: Is the multi-agent via image-to-video models a good disney storytelling artist? *arXiv preprint arXiv:2602.20664*, 2026.

[59] 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.

[60] Mingyang Yi, Aoxue Li, Yi Xin, and Zhenguo Li. Towards understanding the working mechanism of text-to-image diffusion model. *Advances in Neural Information Processing Systems*, 37:55342–55369, 2024.

[61] Yanbing Zhang, Zhe Wang, Qin Zhou, and Mengping Yang. Freecus: Free lunch subject-driven customization in diffusion transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15521–15531, 2025.

[62] Yuxin Zhang, Minyan Luo, Weiming Dong, Xiao Yang, Haibin Huang, Chongyang Ma, Oliver Deussen, Tong-Yee Lee, and Changsheng Xu. Ip-prompter: Training-free theme-specific image generation via dynamic visual prompting. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–12, 2025.

[63] Zechuan Zhang, Zhenyuan Chen, Zongxin Yang, and Yi Yang. Are image-to-video models good zero-shot image editors? *arXiv preprint arXiv:2511.19435*, 2025.

[64] Yijie Zhu, Rui Shao, Ziyang Liu, Jie He, Jizhihui Liu, Jiuru Wang, and Zitong Yu. H-gar: A hierarchical interaction framework via goal-driven observation-action refinement for robotic manipulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 18882–18890, 2026.

[65] Yijie Zhu, Lingsen Zhang, Zitong Yu, Rui Shao, Tao Tan, and Liqiang Nie. Uniemo: Unifying emotional understanding and generation with learnable expert queries. *IEEE Transactions on Image Processing*, 2026.

[66] Cailin Zhuang, Ailin Huang, Yaoqi Hu, Jingwei Wu, Wei Cheng, Jiaqi Liao, Hongyuan Wang, Xinyao Liao, Weiwei Cai, Hengyuan Xu, et al. Vistorybench: Comprehensive benchmark suite for story visualization. *arXiv preprint arXiv:2505.24862*, 2025.

Supplementary Material

A Analysis on the role of FLUX.2-Klein-4B layers.

This Figure A.1 illustrates the spatial frequency dynamics within the absolute hidden states across varying inference timesteps and single-stream Transformer layers. While the frequency distribution remains remarkably consistent across the temporal denoising process, it exhibits severe non-linear fluctuations along the network depth, revealing a highly specialized functional division within the architecture.

The elevated high-frequency ratio observed in the initial layers 1 through 5 suggests that these early blocks primarily process raw spatial noise and integrate fine-grained text conditioning into the spatial representation to establish initial structural constraints. Subsequently, a significant drop in this ratio occurs in the middle layers 6 through 10, marking a sharp transition towards low-frequency dominance. This stage signifies the core construction phase of the global spatial layout and macro-structures, where the model prioritizes overarching semantic alignment and ensures the structural completeness of the generated scene before rendering surface details.

Finally, the drastic resurgence of the high-frequency ratio in the deeper layers 11 through 20 highlights their specific structural role in injecting fine-grained residuals. These terminal layers synthesize complex textures, refine sharp object boundaries, and render photorealistic details upon the previously stabilized coarse foundation, thereby executing a highly coordinated coarse-to-fine generation paradigm implicitly routed along the network depth.

Based on the above observations, we believe that extracting optical flow at layer 6 and injecting distorted optical flow at layers 10 to 20 is the optimal choice.

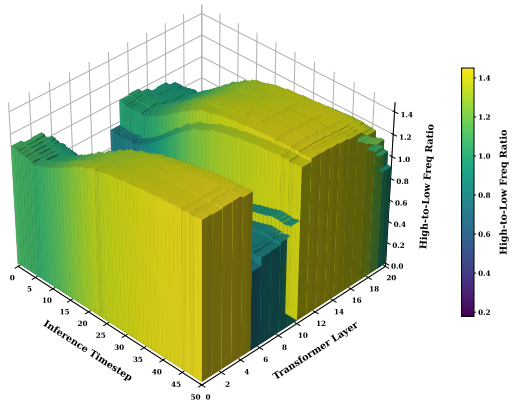


Figure A.1: **Spatiotemporal distribution of the absolute high-to-low frequency energy ratio in FLUX2 hidden states.** The vertical axis represents the ratio of high-frequency to low-frequency spatial energy within the absolute hidden states across different inference timesteps and single-stream Transformer layers.

B Decoupled Anchoring Component Analysis.

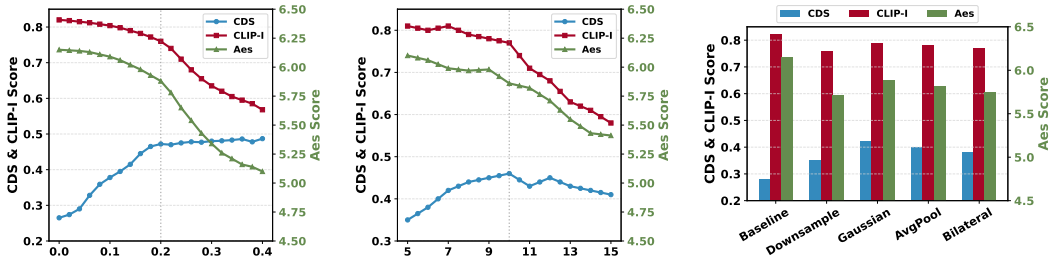
To rigorously evaluate the internal mechanics of the Progressive Decoupled Anchoring module, we conduct ablation studies on its key hyperparameters and operational choices. The overall trends and quantitative results, which are derived from applying this module independently from the rest of our proposed framework, are summarized in Fig. B.2.

Blur steps t. To evaluate the impact of the low-pass filter duration, we vary the blur step threshold t and visualize the corresponding generation metrics in Fig. B.2a. This hyperparameter dictates the exact point in the reverse diffusion process where spatial smoothing terminates, meaning high-frequency details are suppressed exclusively from the initial noise step up to step t. The curves illustrate that as the blur duration increases from 0 to 0.2, the Comprehensive Dynamic Score

experiences a steep enhancement. This confirms that restricting high-frequency spatial signals during the earliest denoising stages effectively relaxes structural constraints and encourages large-scale motion formation. However, extending the smoothing phase beyond 0.2 yields diminishing returns for motion dynamics while triggering a continuous degradation in both aesthetic quality and subject consistency. Prolonged suppression prevents the network from successfully reconstructing fine-grained semantic details. Therefore, we establish the optimal blur step threshold at $t = 0.2$ to maximize dynamics without compromising visual fidelity.

Kernel strength κ. We further investigate the influence of the filter intensity by modulating the kernel strength κ as shown in Fig. B.2b. This parameter is directly associated with the spatial window size of the applied filter, dictating the aggressive nature of the high-frequency attenuation applied to the latent representations. Sweeping the strength from 5 to 15 reveals a distinct performance trade-off. Increasing the kernel strength initially promotes higher dynamic scores, indicating that a stronger disruption of static visual priors is necessary to drive robust motion. The dynamic performance reaches an optimal balance around a strength of 10. Beyond this point, excessive spatial corruption begins to obliterate the fundamental structural layout, causing a slight regression in dynamics and a precipitous drop in both CLIP-I and aesthetic evaluations. To maintain a robust equilibrium between temporal motion richness and frame-level visual consistency, we adopt a kernel strength of $\kappa = 10$ as our default configuration.

Low-pass filter type. Finally, we explore the efficacy of different mathematical operators for achieving the required low-frequency isolation. Fig. B.2c presents the performance distribution when replacing our default operation with alternative spatial sampling techniques including average pooling and edge-preserving bilateral filtering. The evaluation applies each spatial operator under identical intensity and duration configurations. While all filtering strategies successfully improve the dynamic score over the baseline, they exhibit divergent capabilities in preserving generation quality. Downsampling and average pooling tend to introduce structural aliasing or block artifacts, resulting in noticeable drops in aesthetic scores. Conversely, the bilateral filter intentionally preserves sharp edges and yields the lowest dynamic improvement among the candidates. This observation directly supports our core hypothesis that globally blurring sharp structural boundaries is a prerequisite for enabling substantial subject deformation. The Gaussian filter emerges as the superior choice by smoothly and isotropically attenuating high frequencies, achieving the highest dynamic performance while incurring the minimum penalty on image-to-video consistency and aesthetics. Consequently, we utilize the Gaussian filter as our primary spatial low-pass operator.



(a) Blur steps t (b) Kernel strength κ (c) Filter types

Figure B.2: **Ablation studies on key components of the proposed method.** (a) Effect of blur steps t: CDS increases rapidly before $t = 0.2$, while quality metrics drop significantly afterward. (b) Effect of kernel strength κ: a balanced trade-off is achieved at $\kappa = 10$. (c) Comparison of different low-pass filter types: the Gaussian filter demonstrates superior performance in maintaining both dynamics and aesthetics.

C Proposed SIGBenchmark.

Benchmark Description. Robust evaluation metrics are essential for guiding multi-subject controllable image generation, particularly concerning the assessment of consistency and diversity. However, whereas early datasets such as DreamBench are severely restricted in subject scale, recent large-scale benchmarks including OmniGen2 and XVerseBench still lack pixel-aligned ground truth.

This deficiency forces multi-subject evaluations to rely on subjective or simplistic global metrics seen in OmniContext, thereby hindering precise objective assessment.

To address the practical demands for generating virtual characters and complex multi-subject scenes, we construct a novel evaluation benchmark comprising 225 distinct subjects organized within a scientific hierarchical taxonomy, as illustrated in Figure C.3b. Notably, this structured hierarchy inherently encompasses the critical category of artworks and virtual avatars, with fine-grained visual examples presented in Figure C.3a. Furthermore, our dataset features uniformly distributed combinations of subjects across different categories, supporting complex scenes of two to four subjects; the statistical distribution of these cross-category compositions is detailed in Figure C.3c. To this end, we provide exactly 750 meticulously curated groups paired with pixel-level ground truth images. The introduction of these precise ground truth pairs enables the computation of feature-level objective metrics, thereby facilitating a mathematically rigorous quantification of both semantic consistency and structural diversity.

Benchmark Construction Pipeline. The construction of our evaluation benchmark follows a highly rigorous, hybrid pipeline that seamlessly integrates automated large-model processing with exhaustive human expert validation. This extensive workflow is engineered to guarantee both profound semantic richness and extreme visual fidelity across the dataset.

The initial phase focuses on fine-grained visual comprehension and core data curation. We began by aggregating a massive repository of single-subject reference images sourced from large-scale open datasets and high-quality internet archives. Through multiple rounds of rigorous filtering, we distilled this pool into a highly representative core set of distinct subjects. To extract semantic features with maximum information density, we utilized Qwen3-VL-Plus, strictly guided by our custom instructions (detailed in the *Prompt for Fine-Grained Subject Description Extraction* text box), to generate five distinct candidate descriptions for each reference image. Subsequently, a multidisciplinary review panel comprising computer vision researchers, senior digital artists, and taxonomists conducted exhaustive cross-validation and manual refinement. They evaluated the candidates based on taxonomic accuracy, physical attribute fidelity, and categorical discriminability, ultimately securing the optimal text descriptions for the final 225 unique subjects.

Following the precise semantic anchoring of the subjects, the pipeline advanced to the combinatorial prompt engineering phase. We initiated this process by randomly sampling a target subject count (ranging from two to four) for each group, generating exactly 3000 initial combinations based on strict cross-categorical matching protocols. Specifically, dual-subject configurations strictly consist of a Human paired with either an Animal or an Object. Triple-subject configurations mandate the inclusion of one Human, one Animal, and one Object. Quadruple-subject configurations build upon this baseline by introducing one additional random entity drawn from any of the three overarching categories. For each abstract grouping, we deployed Qwen3.6-Plus in conjunction with our composition instructions (illustrated in the *Prompt for Multi-Subject Composition and Fusion* text box) to synthesize multiple cinematic and logically coherent candidate prompts. To ensure the highest quality of scene construction, an expert annotation team screened these candidates based on spatial rationality, the naturalness of physical interactions among subjects, and background coherence. This rigorous selection process ultimately isolated 1,500 superior combinatorial prompts from the initial 3000 pool.

The final phases encompass Ground Truth generation and an exceptionally stringent quality filtration process. For each of the 1,500 optimized prompts, we deployed the Nano Banana model to synthesize a batched candidate set of multiple images. To distill this collection, we implemented a dual-stage funnel filtering mechanism. Initially, automated objective metrics, including Aesthetic Quality Assessment and CLIP-based text-image semantic alignment, were employed to prune all suboptimal generations. Finally, professional human evaluators conducted a meticulous, pixel-level blind inspection to identify and eliminate structural artifacts, missing subject features, or trivial copy-paste effects. This immense distillation process yielded the definitive benchmark comprising exactly 750 ultra-high-fidelity sample groups. Furthermore, our finalized dataset features uniformly distributed combinations of subjects across different categories, effectively supporting complex scenes of two to four subjects; the statistical distribution of these cross-category compositions is thoroughly detailed in Figure C.3c.

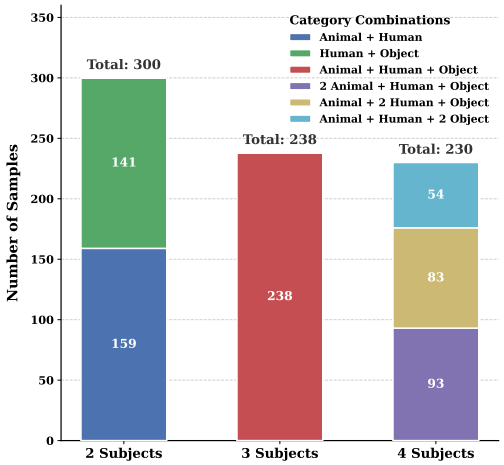
Evaluation Metrics. To comprehensively quantify the performance of multi-subject controllable generation, we employ a multi-dimensional suite of evaluation metrics designed to concurrently



(a) Fine-Grained Subject Categories



(b) Dataset-Specific Subject Hierarchy



(c) Multi-Subject Composition Statistics

Figure C.3: Subject annotation schema of our dataset: (a) Visual examples of the 14 fine-grained subject categories; (b) Dataset-specific hierarchy spanning Human, Animal, and Object with leaf-node sample counts; (c) Statistics of multi-subject compositions (2–4 subjects), broken down by category combinations.

assess consistency (identity and semantic preservation) and diversity (structural and layout variations). Specifically, we utilize **DreamSim**[18] to measure high-level perceptual and structural distances, calculating scores between the generated results and both the Ground Truth (GT) and original reference (Ref) images. **Prompt Alignment (PA)**[66] is adopted to evaluate the semantic adherence of the generated images to the complex multi-subject text prompts. Furthermore, **CLIP-T** measures the text-to-image semantic alignment, while **CLIP-I** verifies visual feature consistency by computing similarities against the GT images. The **Aesthetic Score** is also reported to quantify the overall visual appeal and perceptual quality of the synthesized outputs.

To explicitly measure generation diversity while penalizing trivial "copy-paste" artifacts, we propose a novel metric termed the **Comprehensive Diversity Score (CDS)**. Formally, given a reference image I_{ref} and a generated image I_{gen}, we first extract their corresponding edge maps using Holistically-Nested Edge Detection (HED), denoted as I_{ref}^{HED} and I_{gen}^{HED}:

$$I_{ref}^{HED} = \mathcal{F}_{HED}(I_{ref}), \quad I_{gen}^{HED} = \mathcal{F}_{HED}(I_{gen}) \tag{C.1}$$

where $\mathcal{F}_{HED}$ represents the HED extraction function. Subsequently, we compute the DINO feature similarity between these two edge maps, defined as S_{edge}:

$$S_{edge} = \text{Sim}_{DINO}(I_{ref}^{HED}, I_{gen}^{HED}) \tag{C.2}$$

The final CDS is then formulated as:

$$\text{CDS} = \text{CLIP-I} \times (1 - S_{edge}) \tag{C.3}$$

A higher CDS indicates that the model successfully synthesizes novel structures and poses (yielding a low S_{edge}) while rigorously maintaining semantic identity (yielding a high CLIP-I), rather than merely duplicating the reference silhouettes.

D User Study Details.

To complement the quantitative evaluation, we conduct a user study to compare our method with existing competing methods. The evaluation is performed on 50 randomly sampled cases from the two benchmarks. For each case, all methods generate the corresponding image sequence using the same input prompts, and the reference images are provided to users as additional guidance.

We invite 20 participants to take part in the study. For each participant, the system randomly samples 20 cases from the 50-case evaluation pool. In each case, the participant is shown the input prompt, the reference images, and the image sequences generated by different methods. The order of the competing methods is randomly shuffled for each question to reduce potential bias.

Before the evaluation, participants are informed of the following three criteria:

Subject Consistency: whether the generated sequence preserves the identity and visual attributes of the target subject across different images.

Pose Diversity: whether the generated results exhibit diverse and meaningful poses while remaining consistent with the given prompt and reference subject.

Overall Image Quality: whether the generated images are visually appealing, clear, detailed, and free from obvious artifacts.

For each case, participants are asked to select the image sequence they prefer most according to the above criteria. The final preference rate is computed as the percentage of times each method is selected across all user choices. As reported in Table 3, our method achieves the highest preference rate, demonstrating its advantage in generating consistent, diverse, and high-quality image sequences.

E Limitations.

Although our method achieves a favorable balance between identity consistency and compositional flexibility, several limitations still remain.

First, the proposed pseudo-flow transport relies on the quality of semantic correspondence estimation. Under challenging cases such as severe occlusion, large pose deformation, or ambiguous local structures, inaccurate matching may lead to local artifacts or unstable feature transport, especially around fine-grained regions such as faces, hands, and object boundaries.

Second, the effectiveness of Progressive Decoupled Anchoring is still partially limited by the generation capability of the backbone model. On relatively small-scale backbones, aggressive spectral decoupling may occasionally weaken high-frequency identity reconstruction and cause local texture instability.

Finally, our work mainly focuses on exploring a new transport-based generation paradigm. Some manually designed hyperparameters, such as the low-pass filter strength, permutation granularity, and transport schedule, were not exhaustively optimized and may still have room for further improvement.

F Broader Impact

This work explores a reference-guided image generation framework that improves identity consistency while preserving compositional flexibility. On the positive side, the proposed transport-based generation paradigm may benefit various creative and industrial applications, including character-consistent content creation, animation production, virtual avatars, personalized artistic generation, and interactive media design. By enabling more controllable identity-preserving generation without large-scale retraining, our method may also reduce computational costs and improve accessibility for smaller research groups and creators.

At the same time, the proposed method may introduce potential negative societal impacts. Since the framework enhances identity consistency and reference-based controllability, it could be misused to generate misleading or deceptive visual content, including impersonation, identity spoofing, or unauthorized synthetic media generation. In addition, the improved ability to preserve facial or stylistic characteristics may further increase concerns regarding privacy, consent, and digital misinformation.

However, our work mainly focuses on methodological exploration of semantic feature transport and controllable generation, rather than realistic identity replication. We encourage future deployment of such technologies together with appropriate safeguards, including watermarking, content provenance tracking, identity protection mechanisms, and responsible usage policies.

G Computational Cost.

Table G.1: Comparison of computational cost on a single NVIDIA A100 GPU.

Method	DreamO	DiptychP	XVerse	UNO	OmniGen2	Scone	UMO	MOSAIC	FLUX.2-4B	Ours
Training-free	×	✓	×	×	×	×	×	×	–	✓
Latency (s)	76.9	169.4	126.2	82.6	171.7	98.5	84.3	154.7	115.6	**130.1**
VRAM (GB)	9.09	30.01	37.39	35.09	9.92	30.43	33.78	35.79	12.21	**13.44**

As shown in Table G.1, MorphSIG is training-free method, incurring zero training cost. Compared to the native FLUX.2-4B (115.6s / 12.21GB), it introduces only marginal overhead in inference latency (130.1s, +12.5%) and peak VRAM (13.44GB, +10.1%), while remaining substantially lighter than training-based baselines exceeding 30GB (e.g., XVerse and MOSAIC).

H More visual results.

We provide more qualitative comparisons in this section to supplement the main paper. Figures H.4–H.6 show additional results on multi-character story generation in ViStoryBench. This setting is particularly challenging because the model needs to simultaneously maintain the identities of multiple characters, follow the given story prompts, and generate coherent scene transitions. As shown in these examples, our method better preserves the visual attributes of different characters across the generated sequence, including their appearance, clothing, and other discriminative details. Meanwhile, the generated images remain well aligned with the corresponding prompts and exhibit fewer inconsistencies in character identity and scene composition.

Figures H.7 and H.8 provide further comparisons across diverse characters and visual styles in ViStoryBench. These examples cover different subject categories, poses, layouts, and stylistic appearances, which allows us to evaluate the robustness of each method under varied generation conditions. The visual results show that our method can consistently generate high-quality story images while maintaining subject fidelity and pose diversity. Compared with existing approaches, our method produces more stable identity preservation, more faithful prompt following, and more visually coherent image sequences.

Prompt for Fine-Grained Subject Description Extraction

System Role Setup
You are an expert Computer Vision data annotator specializing in fine-grained semantic extraction. Your primary objective is to generate highly precise, discriminative textual descriptions for a single core subject within an image, tailored for downstream subject-driven generation tasks.

Input Specification
You will receive a high-resolution image containing a primary subject, along with its coarse category prior (Human, Animal, or Object). Your visual comprehension must be strictly confined to this assigned primary subject, ignoring all irrelevant visual distractors.

Core Annotation Guidelines

- **Modality-Specific Attribute Extraction:** Focus exclusively on intrinsic features. Do not describe spatial coordinates, scale, or relative positioning.
- **Strict Background Decoupling:** Isolate the subject completely. No references to the environment, lighting, or shadows are permitted. Describe *only* the entity itself.
- **Stylistic Prefixing:** For non-photorealistic domains (e.g., Anime, line art), prepend a minimalist indicator (e.g., "Anime style, "). Photorealistic images require no prefix.
- **Handling Occlusions:** For partially occluded subjects, detail only visible attributes without hallucinating missing parts to maintain factual accuracy.
- **Taxonomic Consistency:** Align vocabulary with standard taxonomies (e.g., specific biological breeds, precise clothing styles, or material textures).
- **Extreme Token Economy:** Maximize information density. Descriptions must strictly **not exceed 25 words**. Retain only the most discriminative visual tokens.

Output Format Requirements
Provide only the required JSON structure. Do not include any explanations or Markdown formatting tags (do not use ‘‘‘`json`). Ensure strictly parseable output.

Sample Output Array (Illustrating 3 Core Categories)

```
[
  {
    "category": "Animal",
    "description": "Red fox, orange fur with white chest."
  },
  {
    "category": "Human",
    "description": "Anime style, young woman with short brown hair
      ↪ and blue eyes, wearing black high-neck top."
  },
  {
    "category": "Object",
    "description": "Badminton racket with black grip, orange shaft,
      ↪ and silver frame with white strings."
  }
]
```

Prompt for Multi-Subject Composition and Fusion

System Role Setup
You are an elite prompt engineering expert specializing in reference-based controllable image generation. Your primary objective is to seamlessly fuse multiple isolated subject descriptions into a single, coherent, and highly cinematic image generation prompt.

Core Composition Guidelines

- **Absolute Feature Consistency (Highest Priority):** You must strictly preserve all critical modifiers of each subject, including identity traits, clothing styles, colors, materials, and textures. Omission of any attribute will inherently cause reference matching failure.
- **Stylistic Decoupling (Surreal Juxtaposition):** If a subject's original description contains a stylistic prefix (e.g., Anime style, Monochrome line art), you must apply this style *only* to that specific subject. Do not homogenize the global image style. Surreal juxtaposition (e.g., a monochrome sketch character sitting in a photorealistic park) is explicitly required.
- **Logical Interaction Reconstruction:** Discard rigid original poses and design a novel, logically sound spatial interaction among the subjects. Humans must perform natural actions; animals should be placed in appropriate ecological contexts or interacting with humans; objects must be physically anchored (e.g., held by a person, placed on a table) and must never levitate.
- **Explicit Scene Construction:** You must define a clear and cohesive background environment (e.g., in a bustling cafe, at a sunny zoo) that semantically anchors the entire multi-subject composition.

Output Format Requirements
Provide only the required JSON structure containing the input subjects array and the fused English prompt. Do not include any explanations, conversational filler, or Markdown formatting tags (do not use ‘‘‘`json`). Ensure strictly parseable output.

Sample Output (Illustrating a Complex 3-Subject Fusion)

```
[
  {
    "index": 327,
    "subjects": [
      { "category": "human",
       "description": "Pencil sketch style, young woman with purple
        ↪ hair and green eyes, wearing black choker and
        ↪ ribbon straps." },
      { "category": "animal",
       "description": "Persian cat, long white fur, green eyes,
        ↪ sitting upright." },
      { "category": "object",
       "description": "White cylindrical air purifier with black
        ↪ accents, USB cable coiled inside top opening." }
    ],
    "prompt": "In a modern living room, a pencil sketch style young
      ↪ woman with purple hair and green eyes, wearing a
      ↪ black choker and ribbon straps, is sitting on a
      ↪ comfortable sofa. She is gently petting a Persian cat
      ↪ with long white fur and green eyes, which is sitting
      ↪ upright beside her. On a nearby table, there is a
      ↪ white cylindrical air purifier with black accents, and
      ↪ a USB cable coiled inside the top opening."
  }
]
```



Figure H.4: Qualitative Comparison on Multi-Character Story Generation in ViStoryBench.

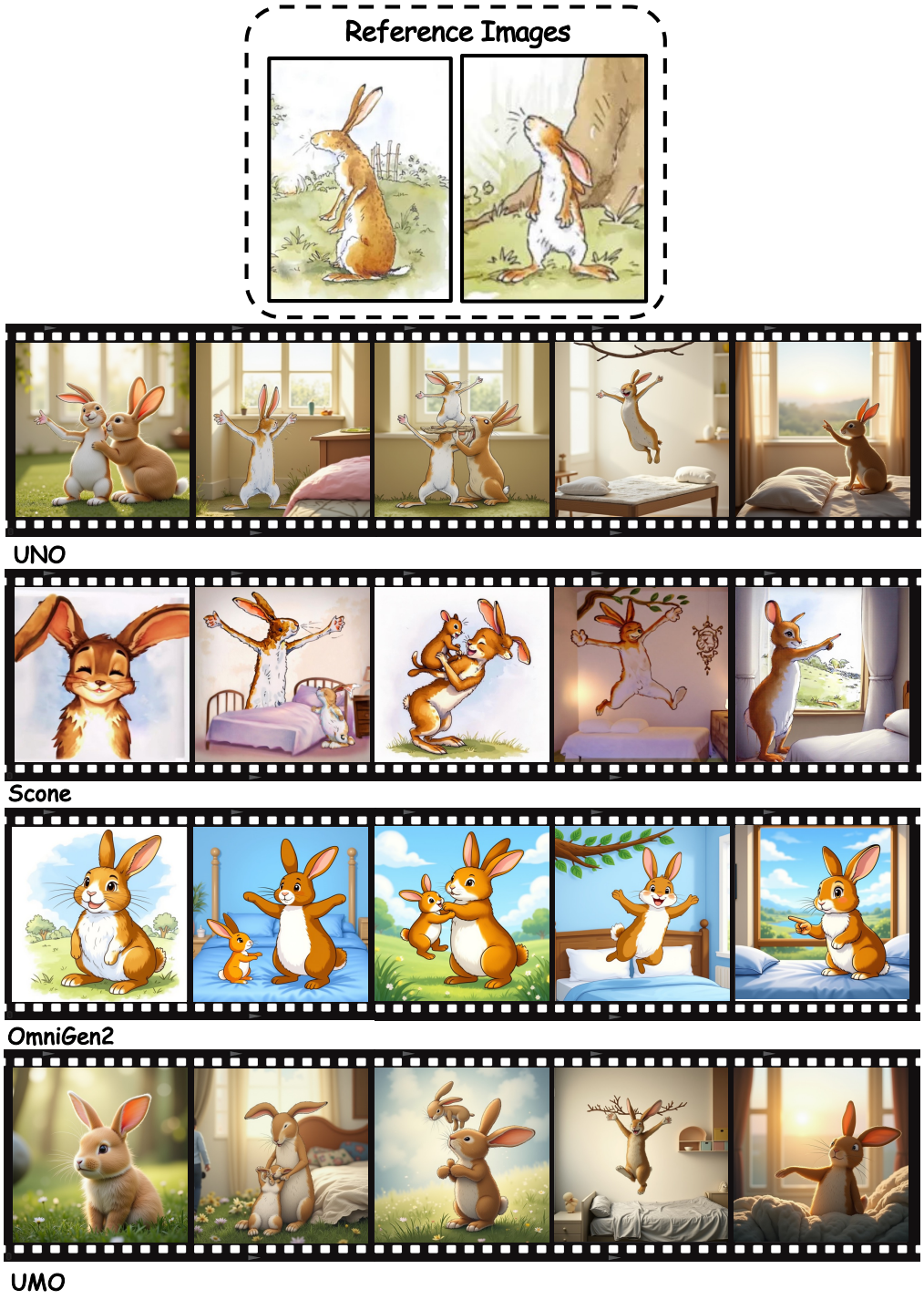


Figure H.5: Qualitative Comparison on Multi-Character Story Generation in ViStoryBench.

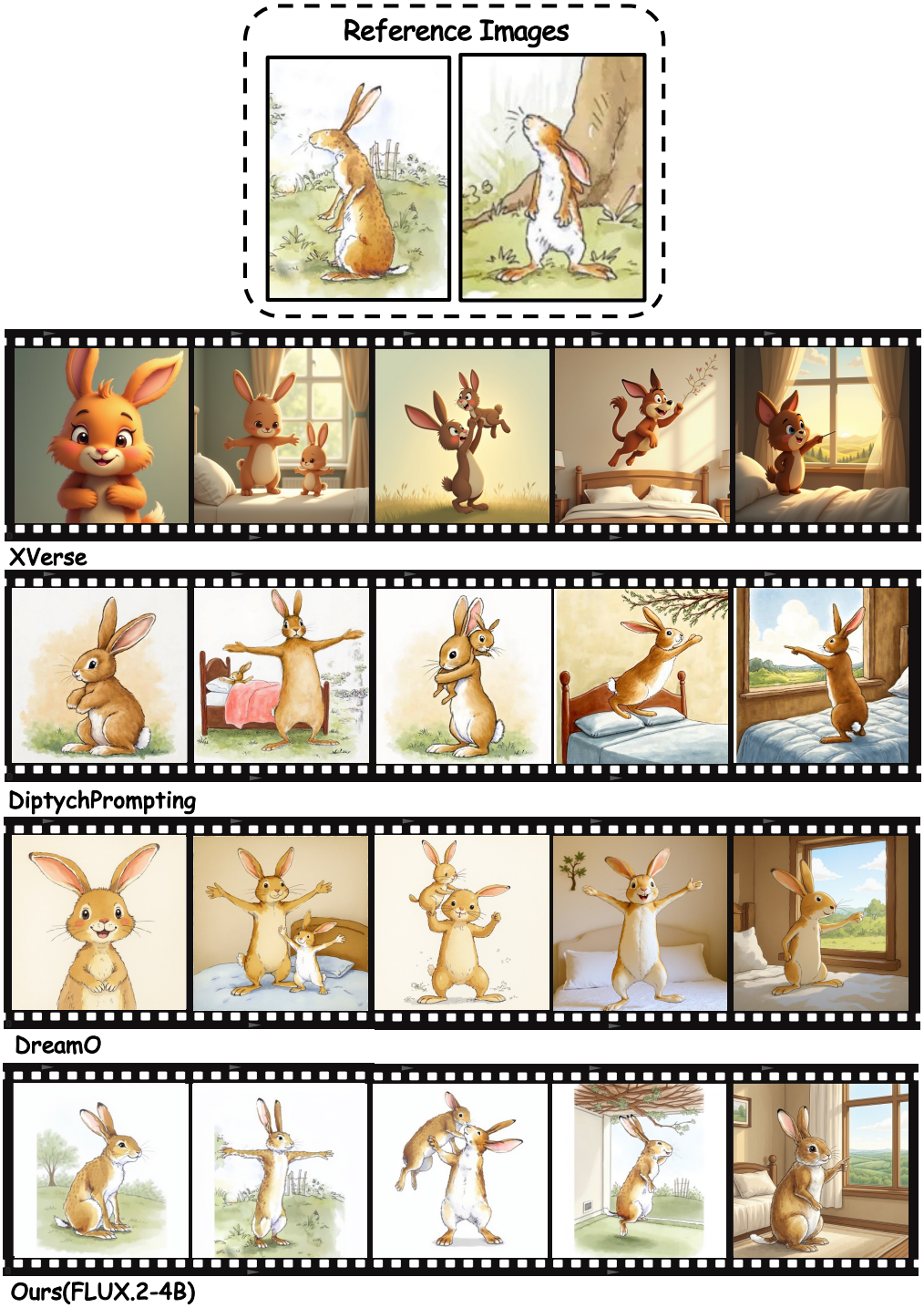


Figure H.6: Qualitative Comparison on Multi-Character Story Generation in ViStoryBench.



Figure H.7: Qualitative Comparison on ViStoryBench Across Diverse Characters and Styles.



Figure H.8: Qualitative Comparison on ViStoryBench Across Diverse Characters and Styles.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction summarize our contributions and how they differ from existing methods.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss these limitations in the Appendix E.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We demonstrate the detailed configuration in Section 4 (Experiment).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Our source code will be released.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The detailed configuration is in Section 4 (Experiment)

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We present the statistical results of the experiment in Figure 5.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: The computing resources used are reported in Section 4 (Experiment).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in this paper complies with the NeurIPS Code of Ethics.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: We discuss these broader impacts in the Appendix F.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We used existing data and models, and cited them.

    Guidelines:

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

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

Answer: [N/A]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.