StoryTailor: A Zero-Shot Pipeline for Action-Rich Multi-Subject Visual Narratives

Supplementary Material

Method	VAS-S↑	VDM-S↑	VAS-MS↑	VDM-MS↑
QWEN	1.32	0.73	-2.88	0.19
Nano	1.74	1.25	-1.55	0.87
FluxKontext	0.26	0.09	-2.93	0.12
MS+1P1S	-3.61	-2.73	-6.89	-3.58
Ours	**2.21**	**1.63**	**-0.22**	**0.74**

Table 5. SigLIP-based action scores, where higher values indicate better performance. S denotes single-subject and MS denotes multi-subject settings.

6. Additional Experiments

6.1. Action Boost Evaluation

To provide a more comprehensive evaluation of our method's advantage in action generation, we further introduce SigLIP-based action metrics to perform quantitative analysis from two perspectives: action alignment and action discriminability. Specifically, we first construct a predefined candidate verb set and use SigLIP to compute the matching logits between each generated image and the text of all candidate verbs. Based on these scores, the Verb Alignment Score (VAS) is defined as the SigLIP logit of the ground-truth action verb within the predefined verb set, measuring how strongly the generated result responds to the target action semantics. The Verb Discriminability Margin (VDM) is defined as the difference between the logit of the ground-truth action verb and the average logit of the Top-10 candidate verbs, measuring the discriminative margin of the target action against competing actions, that is, whether the target action can be more clearly separated from semantically similar alternatives. This design not only evaluates whether the model captures the correct action semantics, but also reflects whether its action representation is sufficiently discriminative. As shown in the Tab.5, the results demonstrate that our method not only aligns more accurately with the target action semantics, but also improves the discriminability of action representations, thereby providing stronger evidence for its superiority in action generation.

6.2. Single-frame Image Consistency(Addition)

To further assess StoryTailor's ability to preserve subject identity and obey action text in isolated frames, we ran additional single-frame consistency experiments on MSBench [29]. MSBench spans diverse multi-subject scenes. Combinations range from one to three subjects, with 100 long-form prompts per combination; each prompt is paired with per-subject reference images and grounding boxes. All images were generated at 1024×1024 using 50 sampling steps on a single RTX 4090 GPU. We compare StoryTailor against eight strong baselines that cover fine-tuning, adapter-based, in-context, and edit paradigms: LoRA[10], DreamBooth[27], IP-Adapter[33], MS-Diffusion[29], FluxKontext[11], λ-Eclipse[20], Qwen-Edit[30], and NanoBanana[3]. Metrics include CLIP-I for identity preservation, CLIP-T for text alignment, and Dino for subject similarity.

Quantitative. Quantitative results appear in Table 6. StoryTailor achieves a CLIP-I score of 0.849 in single subject cases. This outperforms MS-Diffusion at 0.824. The improvement stems from reduced background carryover through GCA. In multi-subject setups our method boosts CLIP-T by up to 15 percent. It reaches 0.414 compared to 0.340 for MS-Diffusion. AB-SVR amplification of action related embeddings drives this gain. It enhances interaction expressiveness without identity confusion. DINO-v2 and M-DINO remains strong. This indicates superior identity preservation.

In summary, StoryTailor achieves the best CLIP-T for both single- and multi-subject because AB-SVR amplifies verb/interaction directions in text embeddings, GCA softly unbinds subject neighborhoods with a two-stage Gaussian to reduce overlap entanglement, and SFC keeps only transferable background cues while suppressing stale history—trading a bit of identity saturation for stronger text/action faithfulness. FluxKontext emphasizes global context and latent reuse, yielding strong identity (CLIP-I) but weaker verb grounding and much higher cost. Qwen-Edit (API editing) is strong on single-subject DINO-v2/CLIP-T, yet lacks structural cross-frame/ cross-subject constraints in crowded scenes, leading to identity–action trade-offs. Nano-Banana attains top multi-subject CLIP-I/M-DINO via tighter reference binding but remains conservative on verbs, hence lower CLIP-T. MS-Diffusion / SSR-Encoder / λ-ECLIPSE stabilize appearance similarity without explicit action boosting or overlap disentangling, so CLIP-T lags. IP-Adapter is weakest on verbs without added mechanisms. Fine-tuning methods (Textual Inversion/LoRA/DreamBooth) overfit limited views—identity is decent but actions under-expressed and they lack our/FluxKontext-style cross-frame memory and practical efficiency. Overall, the complementarity of GCA + AB-SVR + SFC prioritizes action faithfulness without freezing dynamics, while keeping identity and cost accept-

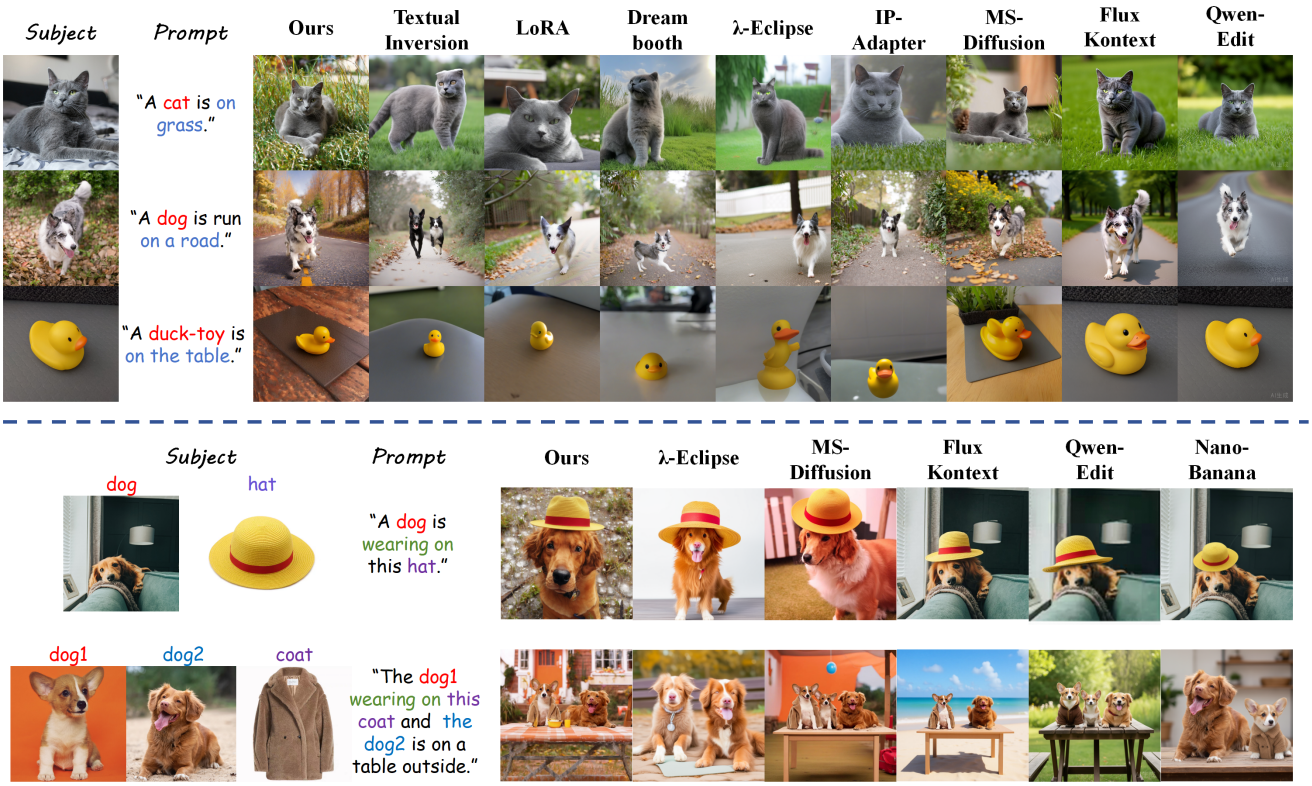


Figure 6. Additional single-frame image consistency visual experiments.

Table 6. Single-frame Image Consistency on MSBench (micro-averaged means). Single-subject reports CLIP-I, DINO-v2, CLIP-T; Multi-subject reports CLIP-I, M-DINO, CLIP-T. **Best values are bold** and second-best values are underlined for *each column*. Inference Time and Peak GPU VRAM are overall efficiency metrics.

	Single-subject			Multi-subject			Efficiency	
Method	**CLIP-I ↑**	**DINO-v2 ↑**	**CLIP-T ↑**	**CLIP-I ↑**	**M-DINO ↑**	**CLIP-T ↑**	**Time (s) ↓**	**VRAM (GB) ↓**
Fine-tuning								
Textual Inversion	0.772	0.574	0.264	–	–	–	3.51	13.73
LoRA	0.795	0.591	0.270	–	–	–	3.85	15.82
DreamBooth	0.802	0.589	0.309	–	–	–	4.25	13.12
Adapters & MLLM								
IP-Adapter	0.809	0.616	0.283	–	–	–	4.82	19.39
MS-Diffusion	0.824	0.848	0.327	0.692	0.108	0.340	5.36	19.47
SSR-Encoder	0.813	0.799	0.331	0.723	0.104	0.311	7.49	18.82
λ-ECLIPSE	0.792	0.806	0.308	0.719	0.097	0.304	8.31	14.21
In-context								
FluxKontext	**0.872**	0.859	0.342	0.732	0.107	0.372	32.31	22.99
Qwen-Edit	0.841	**0.902**	0.373	0.714	0.108	0.396	–	–
NanoBanana	0.857	0.898	0.362	**0.749**	**0.114**	0.389	–	–
StoryTailor (Ours)	0.849	0.811	**0.431**	0.642	0.112	**0.414**	7.02	20.01

able.

Qualitative. We characterize single-frame behavior as the number of subjects increases from one to three on MS-Bench, keeping resolution (1024×1024), steps (30), seeds, boxes, and per-subject references identical across methods; SFC is disabled to isolate single-frame effects. Baselines span LoRA, DreamBooth, IP-Adapter, MS-Diffusion, FluxKontext, λ-Eclipse, Qwen-Edit, and NanoBanana. Results is shown in Fig.6.

6.3. Animals and Inanimate Objects Visual Narrative Task

We illustrate the narrative ability of StoryTailor on both animals and inanimate objects, as shown in Fig. 7. For each subject we provide one reference image and a single long-form prompt that lists a sequence of actions or scene descriptions. The pipeline decomposes the prompt into frame-level clauses and generates a visual story in which every clause corresponds to one image, while the subject identity and style remain stable.

For single-subject animals, the dog sequence covers actions such as "running in forest", "running on the beach by ocean", "standing in rainy day", "sitting on the floor at the owner's feet", and "lying on a bed". StoryTailor preserves the same dog across all frames, while backgrounds, lighting, and camera viewpoints vary in line with the described situations. The transition from forest to beach and finally to indoor scenes remains smooth. No extra dogs appear and the facial markings stay coherent, which shows that Gaussian-Centered Attention keeps the core identity stable while AB-SVR strengthens the verbs that drive the story. For single-subject still life, we use a yellow clock and describe it "by the bed in morning", "on the kitchen counter during breakfast", "on bedside table in night", "on a dusty shelf", and "flopping on the grass". The generated sequence depicts the same clock with consistent material and color under different contexts. The placements of the clock, supporting objects, and environment match the narrative, for example soft bedding, kitchen utensils, cardboard boxes, and outdoor grass. This suggests that our method can also build plausible stories for rigid objects that do not move on their own, by letting the environment and layout carry the narrative.

For multi-subject animals, the dog–cat sequence asks the pair to "run in forest", "rest on the beach by ocean", "nestle on the beach", "lie on the sofa", and "jump onto the table and lie flat". StoryTailor produces a consistent pair whose fur color and body shape remain stable, while their relative poses and distances change in a natural way. The dog and cat share attention rather than collapsing into one hybrid creature, and occlusions such as the cat leaning on the dog are handled cleanly. For multi-subject still life, the coat–trousers sequence describes "hanging up in store", "tried on by a girl", "thrown on the bed in two sets", "hanging in shelf", and "worn by two girls". The results show one coherent outfit across different frames, from display racks to fitting scenes and lifestyle shots. The fabric tone and cut stay aligned with the reference images, while the composition adapts to the described usage scene, for example flat lay on the bed and two-person portrait at the end.

Overall, these examples indicate that StoryTailor can construct readable visual narratives for both animals and static objects under a single prompt. It maintains subject identity, enriches frame-wise actions and layouts, and arranges backgrounds so that the sequence reads as a short story rather than a set of unrelated images.

7. GCA Anisotropic Experiments

7.1. Settings

We perform an anisotropy ablation of GCA on multi-subject action scenes using a unified setup of 1024×1024 resolution, 30 sampling steps, shared seeds, and a single 24 GB GPU. We report micro-averaged CLIP-I and CLIP-T, measure background-drag ratio, log inference latency and peak VRAM, and flag any subject confusion. GCA is kept as the core while the other module is kept the same. To avoid unverified subject confusion and entanglement, we selected non-interactive actions and set the grounding boxes to [[[0.2, 0.2, 0.6, 0.6], [0.45, 0.20, 0.8, 0.60]]], resulting in an overlap region occupying approximately 12% of the total image area, to validate the methods' effectiveness in subject decoupling. We compare the following six strategies:

Box Binary: in-box binary mask with background dummy as a geometric baseline.

XOR Split: pure XOR hard disentangling for overlaps with direct separation in conflict regions.

Gaussian Diffusion Split: pseudo masks from diffusion-time aggregated text attention with Gaussian-strength adaptive splitting over overlaps and map-derived bias injected to image-side logits.

Static Two-Stage: two-stage Gaussian via long-side split and dual centroids with fixed radii that do not vary with semantics.

Single-Stage: single-stage isotropic Gaussian inside the box with soft sharing in overlaps dynamically driven by attention heatmaps.

GCA: two-stage anisotropic Gaussian via long-side split and dual centroids with radii dynamically driven by attention heatmaps.

7.2. Analysis

The Fig.8 contrasts baseline mask methods left: Binary Box, XOR Split, Gaussian Diffusion Split with two-stage variants right: Static Two-Stage, Single-Stage, GCA in generating multi-subject dog-cat scenes across three actions.

Figure 7. Additional animals and inanimate objects visual narrative task experiments.

Table 7. GCA anisotropy ablation on multi-subject action scenes, 1024×1024, 30 steps, shared seed, single 24 GB GPU. Metrics: CLIP-I↑, CLIP-T↑, background-drag↓, subject confusion(Y/N), latency (s)↓, peak VRAM (GB)↓. *Best*=**bold**, *second*=underline.

Method	CLIP-I↑	CLIP-T↑	Bkg Drag↓	Confusion(Y/N)	Latency↓	Peak VRAM↓
Box Binary	0.815	0.285	13%	Y	**6.23**	**20.24**
XOR Split	0.837	0.279	22%	N	6.27	21.51
Gaussian Diffusion Split	0.819	0.313	28%	Y	31.32	22.01
Static Two-Stage	0.878	0.354	11%	N	7.13	20.58
Single-Stage	0.856	0.349	6%	Y	6.96	20.74
GCA	**0.893**	**0.416**	**2%**	N	7.57	20.44

Baselines exhibit identity swaps, static poses, and background inconsistencies e.g., leaked elements in jumping scenes, while GCA produces vivid interactions, accurate actions e.g., dynamic runs and smiles, and stable evolving backgrounds, aligning with the paper's claims of superior overlap resolution and action richness via anisotropic Gaussians. Table. 7 experimental results reveal that binary box masking causes severe identity confusion and background drag in multi-subject scenes, leading to stiff action expressions and lackluster interactions, while Gaussian diffusion splitting offers slight relief from drag but escalates computational demands. Static two-stage variants start to mitigate identity leakage and improve action coherence, yet they are hampered by rigid boundaries that provoke local conflicts. Single-stage dynamic Gaussian outperforms prior strategies, but action attributes remain insufficiently prominent; two-stage dynamic Gaussian GCA excels markedly, producing vivid interactions, precise actions, and gradually stable backgrounds for superior overall narrative fluency. The rationale stems from GCA's anisotropic Gaussians dynamically centering on subject cores to soften overlap boundaries and avert confusion, paired with AB-SVR's reinforcement of action embedding directions to infuse behavioral diversity, and SFC's selective retention of background semantics over frozen history to secure cross-frame continuity without curtailing subject dynamism. This modular synergy deftly resolves the zero-shot pipeline's inherent tensions.

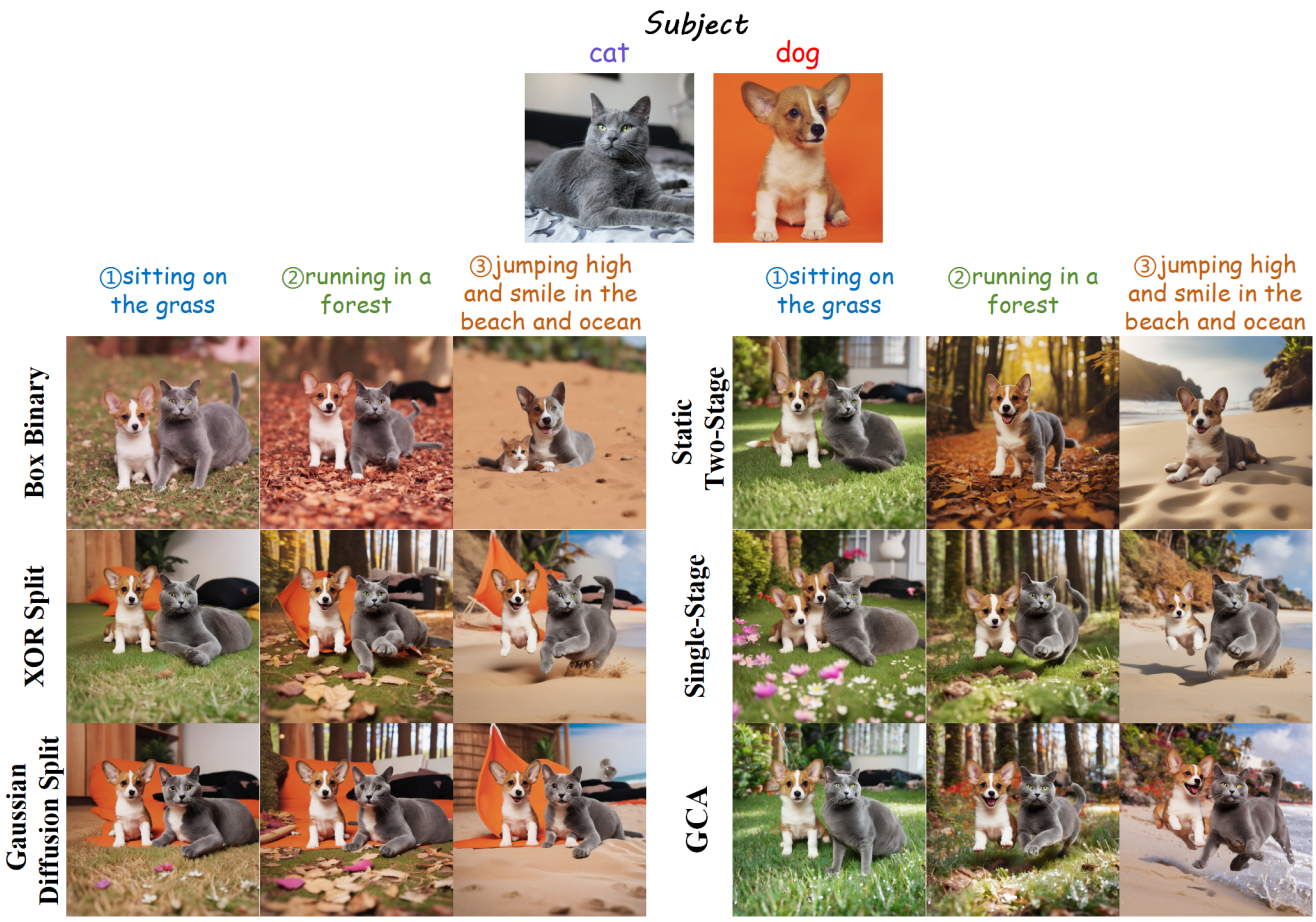


Figure 8. Additional GCA anisotropic comparison experiments.

Table 8. Statistical distribution of knee positions k across cumulative energy thresholds τ in 200 samples.

τ range	$k_{\min}$	$k_{\max}$	k_{avg}	Knees
0.50–0.60	2	3	2.1	18
0.61–0.70	2	4	2.8	32
0.71–0.80	3	6	4.1	41
0.81–0.90	3	8	5.0	53
0.91–0.95	5	18	10.2	37
0.96–0.99	8	39	19.5	19

8. Singluar Value Energy Experiment

To understand how the singular-value energy of each sentence in SVR is distributed over the semantic subspace of the text embedding, we first feed multiple prompts into the text encoder to obtain their embeddings. For each individual sequence of semantic vectors, we perform thin SVD and visualize the decay curves of singular-value energy, which allows us to inspect how different energy bands contribute to the text-embedding subspace. Fig. 9 presents some visual results of singular value decay curves together with statistics of the detected energy knees.

In multi-frame visual narrative generation, we first split the global story prompt into frame-level paragraphs and perform thin SVD on the text embedding of each frame. The frame-wise embedding matrix is factorized along the "feature–token" axes as

$$X = U\Sigma V^T \tag{13}$$

where Σ is a diagonal matrix and each singular value σ_i represents the scaling factor or energy contribution of the i-th principal direction. The squared singular value σ_i^2 corresponds to the variance captured by that component, i.e., the amount of data variability along this direction. Large singular values therefore encode the dominant transformations that carry most of the information in the embedding, whereas small singular values are associated with minor components or noise. At the text-encoding stage, this spectrum induces a natural semantic hierarchy: leading singular values are dominated by subject entities and core at-

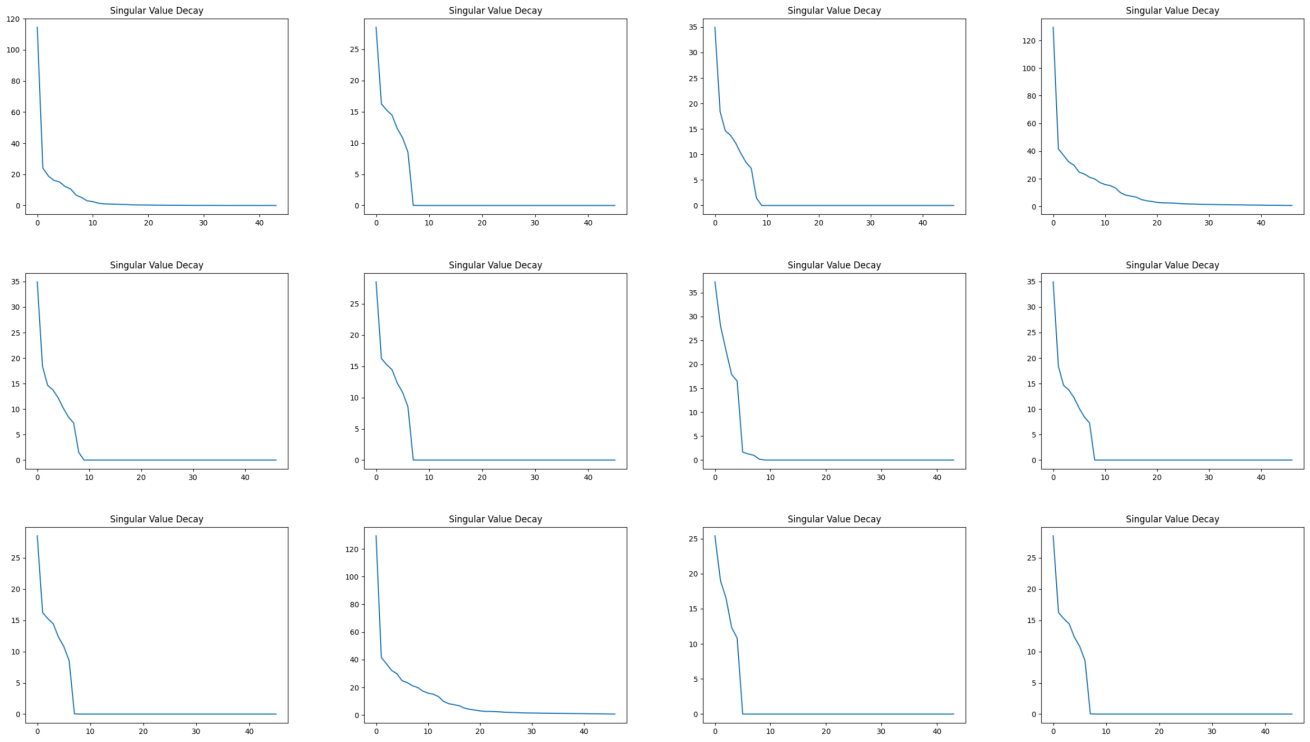


Figure 9. Some visual results of singluar value decay.

tributes, mid-range singular values are more aligned with action details and interaction relations, and tail singular values mostly reflect background style and fine-grained modifiers. A key use of SVD in our setting is low-rank approximation: by truncating smaller singular values, we reconstruct $\mathbf{X}$ with far fewer dimensions while losing only limited information.

We focus on detecting elbows in the singular-value decay curves, because these points mark the boundary between dominant and secondary information. Many real-world matrices, including image features and text embeddings, exhibit a low-rank nature, where the effective rank is much smaller than the nominal dimensionality. In the multi-frame setting, as the number of frames increases, frame-level paragraphs interfere with each other in the embedding space: semantics of non-target frames are suppressed but not completely removed, so residual signals from other frames remain in the principal spectrum. To obtain a cleaner representation for the current frame, we therefore rely on the structure of the singular-value decay. Empirically, the first pronounced energy drop corresponds to backbone subject information, the second major drop aggregates action and interaction semantics, and the third drop mainly captures background and scene style, while the remaining long tail is largely cross-frame residue and noise. Based on this interpretation, we apply low-rank approximation to select singular values within the desired energy bands, retaining subject- and action-related components for the current frame while discarding as much inter-frame noise and redundancy as possible.

As illustrated in Fig. 10, we further visualize the singular-value decay patterns of frame-level text embeddings. For each setting we randomly sample $n \in \{5, 10, 15, 20\}$ frame-level paragraphs, compute their embeddings, and plot the sorted singular values σ_i against the rank k. On every polyline, we mark the three strongest energy drops as the first, second, and third elbows, which approximately correspond to cumulative energy thresholds $\tau \approx 0.60$, $\tau \approx 0.85$, and $\tau \approx 0.95$, respectively. As n increases, these knees consistently cluster into three bands along the k axis, confirming that the spectra of text embeddings exhibit a stable low-rank hierarchy: the first elbow captures subject-centric backbone semantics, the second elbow aggregates action and interaction information, and the third elbow is dominated by background and style. This empirical behavior supports our subsequent choice of τ bands for AB-SVR, where we explicitly operate on these energy regions to retain subject and action components while suppressing residual background and cross-frame noise.

Figure 11 further provides a qualitative view of how the cumulative-energy threshold τ in AB-SVR affects the generated semantics. We fix the subjects (dog or dog+cat)

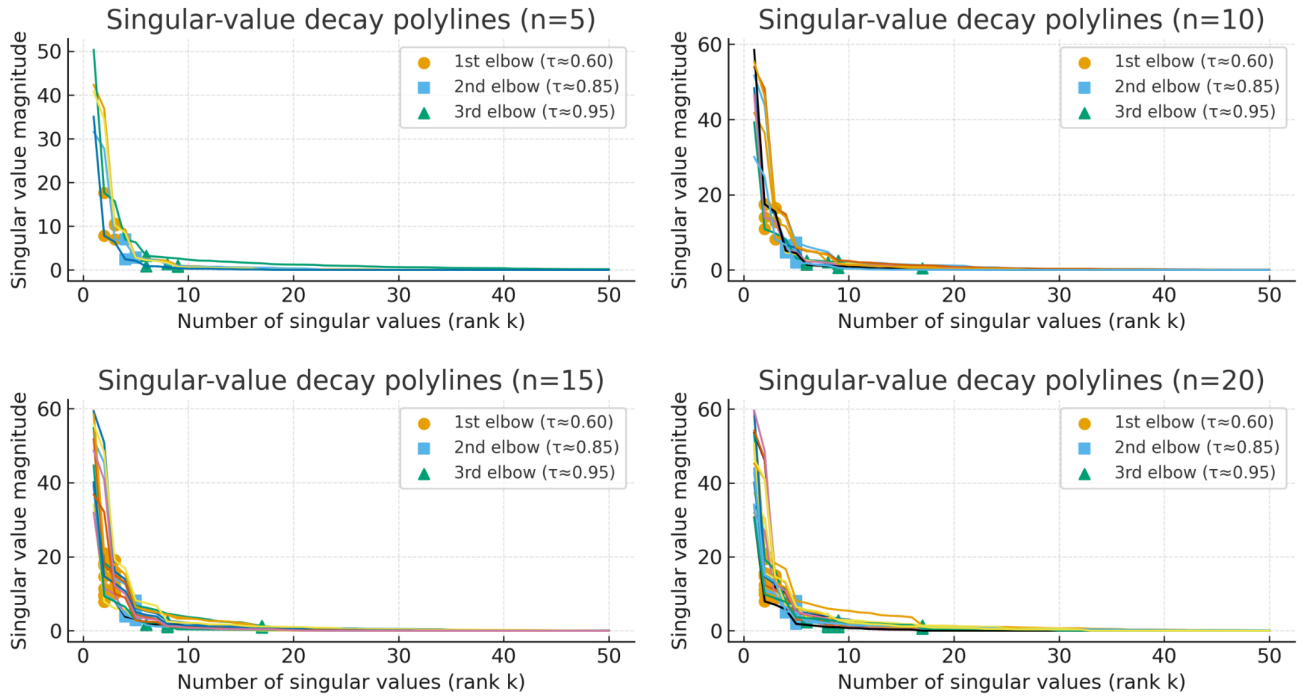


Figure 10. Singular-value decay polylines for different numbers of sampled prompts. Each subplot visualizes $n = 5, 10, 15, 20$ text embeddings, respectively. For every curve we plot the singular-value spectrum and highlight the three largest energy drops with markers: the first elbow (circles, $\tau \approx 0.60$), the second elbow (squares, $\tau \approx 0.85$), and the third elbow (triangles, $\tau \approx 0.95$). The knees cluster into three bands, corresponding to subject-, action-, and background-related energy regions in the text-embedding space.

and frame-level prompts, and sweep τ from 0.50 to 0.98. When τ is too small (e.g., 0.50 or 0.56), only a few leading components are retained: subjects remain roughly identifiable, but actions become attenuated and often collapse into static poses, while backgrounds are under-specified and frequently drift across frames. As τ enters the mid-range band around 0.74–0.80, the reconstructed text embeddings recover more action- and relation-related directions: dogs and cats not only appear with stable identities, but also execute the requested actions (running, jumping, lying) more faithfully, and the surrounding context becomes more coherent without overwhelming the subjects. Pushing τ further towards 0.92 and 0.98 introduces a large portion of background- and style-dominated components, which strengthens texture and scenery but also re-couples subjects and background; in multi- subject scenes this occasionally leads to pose over-regularization or subtle identity blending. These trends are consistent with the spectral analysis in Fig. 10: the first elbow mainly supports subject cores, the second elbow enriches action semantics, and the third elbow starts to absorb background energy, making $\tau \approx 0.80$ a reasonable operating point that balances identity fidelity, action expressiveness, and background consistency.

To better understand how AB-SVR enhances action semantics, we visualize both the long-tail effect in the text embedding spectrum and the attention maps of the text-conditioning branch in Fig. 12. The results show that, after applying AB-SVR, the original long-tail distribution of singular values in the text embeddings is effectively mitigated, and the model produces more concentrated responses to action-related keywords. The attention regions, which were previously scattered over subject appearance or background areas, become progressively focused on key body parts and interaction regions that directly reflect the target action. For example, for actions such as running, jumping, or hugging, the attention maps place greater emphasis on body pose, contact regions, and local structures associated with the action, rather than being dominated by irrelevant scene textures or static appearance cues. These observations suggest that AB-SVR enhances the salience of action-related semantics during cross-attention by reweighting the dominant spectral directions associated with action, thereby improving text-driven action alignment. The visualization results are also consistent with the gains observed in quantitative action-related metrics, further validating the effectiveness of AB-SVR in strengthening action expression.

Figure 11. Singluar value energy τ setting experiments on single- and multi-subject tasks.

9. Extension on DiT Backbone

To verify that StoryTailor is not dependent on a specific U-Net diffusion backbone, we further extend our experiments to DiT-based architectures and replace the original image encoder with SigLIP, in order to evaluate the transferability of our method across different visual representations and generative backbones. Specifically, we choose SD3.5 and FLUX.1 as the underlying backbones, and inject our pipeline using MSAdapter modules retrained for each new backbone, while keeping all other components unchanged. As shown in the Fig. 13, the results demonstrate that, even on DiT architectures that differ substantially from our original implementation, StoryTailor can still preserve coherent action expression and scene continuity in multi-subject visual narratives. These findings suggest that the proposed spatial constraint, spectral reweighting, and selective caching mechanisms are not merely specialized adaptations to the U-Net attention pathway, but instead exhibit a certain degree of backbone-agnostic generalization.

10. Long-frame Narrative Effects

We further study how StoryTailor behaves on long-horizon narratives using the MSBench prompts. For each method, we generate up to 20 frames per story under identical prompts, reference images, bounding boxes, seeds, and sampling steps, and compute DreamSim (↓), CLIP-I (↑), and CLIP-T (↑) for every frame. The curves in Fig. 14–16 are averaged over all sequences.

In terms of DreamSim (Fig. 14), StoryTailor maintains the lowest perceptual distance over almost the entire horizon. Our curve starts around 0.25 and only increases by about 0.03 when moving to 20-frame stories,

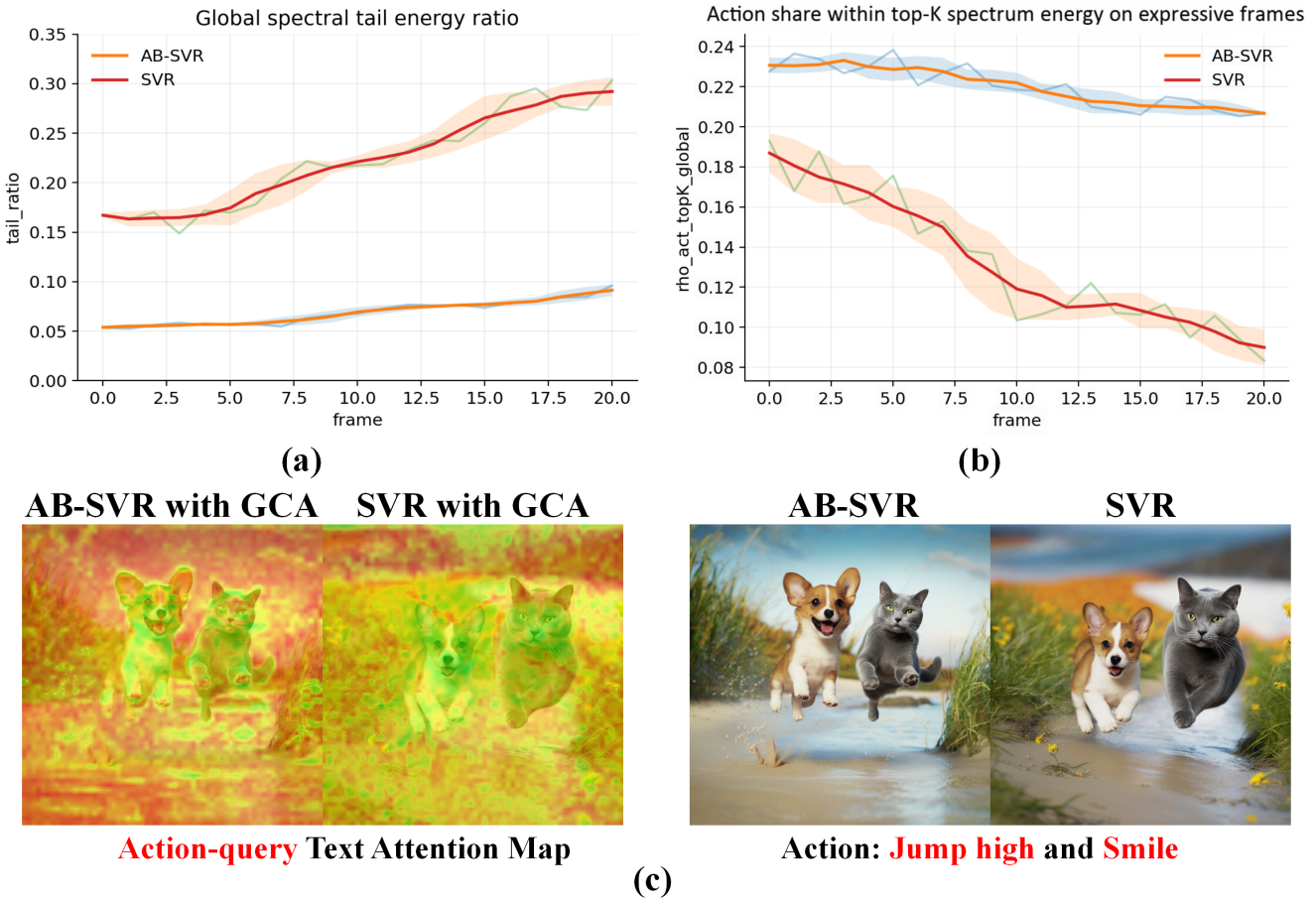


Figure 12. (a)Long-tail effect from high-dimensional accumulation in SVR. (b)Action energy share in singular-value space (AB-SVR vs SVR). (c)Verb query attention heatmap (AB-SVR vs SVR).

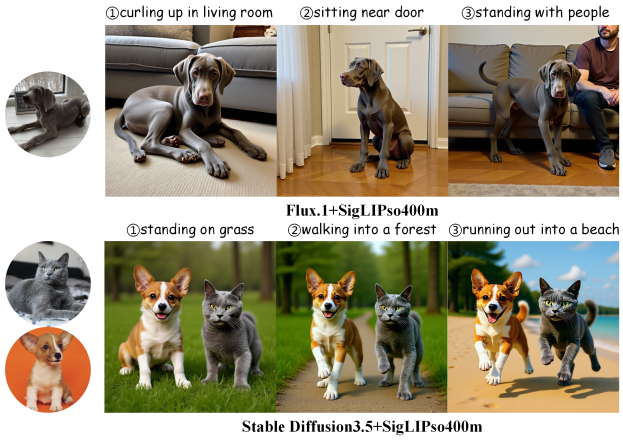


Figure 13. DiT and SigLIP adaptation experiment.

whereas 1p1s rises by roughly 0.16 over the same range and thus accumulates substantial temporal noise. FluxKontext, NanoBanana, and Qwen-Edit stay in the middle: their DreamSim values are more stable than 1p1s but remain consistently higher than StoryTailor at long horizons. This suggests that the combination of Gaussian-Centered Attention, Action-Boost SVR, and Selective Forgetting Cache effectively dampens error accumulation and avoids the "drift into chaos" often observed in long sequences.

CLIP-I results (Fig. 15) show that StoryTailor keeps identity fidelity in a narrow and visually acceptable band around 0.85, with a total variation below 0.01 across all frame indices. In contrast, 1p1s exhibits a clear downward trend and loses about 0.03 in CLIP-I as sequences become longer, indicating gradual identity erosion. In-context baselines (FluxKontext, NanoBanana, Qwen-Edit) retain slightly higher CLIP-I than ours, but qualitative inspection reveals that they do so by producing more static,pose-conservative content. Our method therefore trades a small amount of identity tightness—still acceptable to human observers—for richer and more persistent motion.

For CLIP-T (Fig. 16), StoryTailor consistently achieves the best text–image alignment. Our CLIP-T starts above 0.43 and remains above 0.40 throughout the 20-frame horizon. By comparison, 1p1s drops by almost 0.09, show-

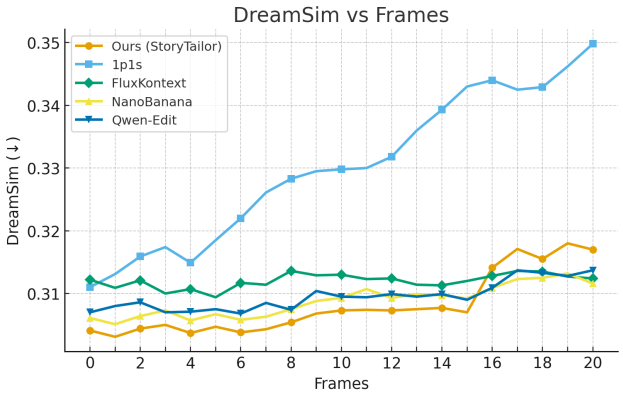


Figure 14. DreamSim vs. sequence length on MSBench (lower is better).

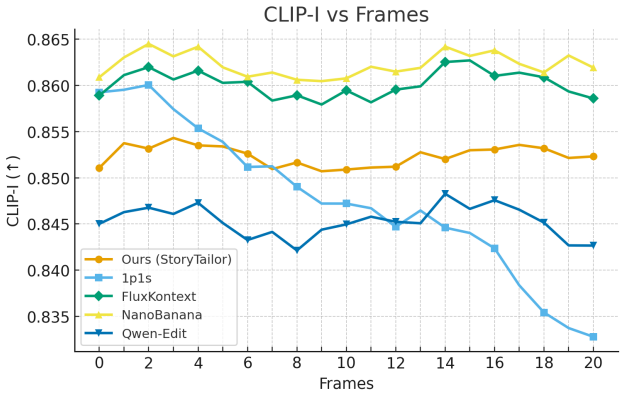


Figure 15. CLIP-I vs. sequence length on MSBench (higher is better).

ing strong long-frame degradation; FluxKontext also decreases notably with length, while NanoBanana and Qwen-Edit decline more gently but stay below StoryTailor for most frames. The relatively flat CLIP-T curve of StoryTailor indicates that Action-Boost SVR repeatedly recenters verb and interaction directions at each frame, and that Selective Forgetting Cache limits the impact of stale history on new textual intents. Overall, these long-frame results confirm that StoryTailor can sustain narrative dynamics over extended horizons: perceptual drift is controlled, identity remains stable, and action semantics stay well aligned with the story text.

Fig. 17 shows a 20-frame single-subject narrative of a dog moving through streets, parks, and seasonal scenes. The dog identity stays highly consistent despite large changes in viewpoint, background, weather, and even style (e.g., snow and black-and-white frames). The sequence covers a diverse set of actions, such as walking, sitting, resting in a sunny spot, digging in the soil, swimming slowly in a pond, and waiting quietly by the door, forming

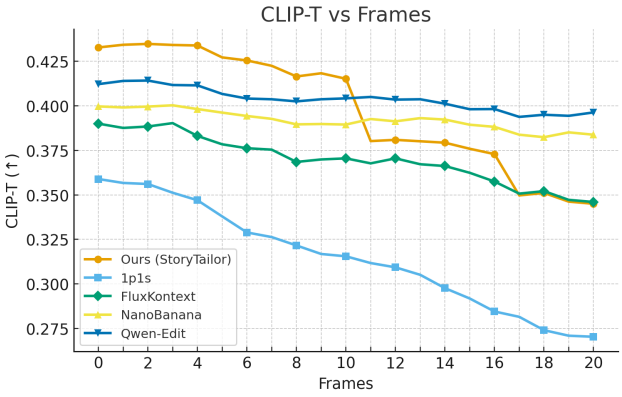


Figure 16. CLIP-T vs. sequence length on MSBench (higher is better).

a plausible progression of everyday activities. Nevertheless, long-frame effects are also visible here: in later frames some prompts with very specific verbs are rendered as more generic "standing" or "sitting" poses, and the distinction between neighboring actions becomes softer than in early frames. Compared with the multi-subject case, identity and layout are easier to preserve, but fine-grained verb expression still tends to be partially realized when the story extends to 20 frames, which again aligns with the slight long-horizon drop in CLIP-T.

Fig. 18 presents a 20-frame narrative with two subjects, a brown dog and a gray cat. The pair traverse the same urban environment while performing a series of explicitly prompted joint actions, including running side by side, sharing a ball, playing tug of war, drinking from the same bowl, and waiting together outside a small shop. Across all frames, both identities remain stable and well separated, the dog–cat layout stays plausible, and the background evolves smoothly along a consistent city route, indicating strong subject fidelity and cross-frame continuity. At the same time, a clear long-frame effect appears in the later part of the sequence: some fine-grained interaction verbs (e.g., "guard the gate," "watch people from the bridge") are only partially realized or simplified into visually similar side-by-side poses, and the difference between consecutive actions becomes less pronounced. This matches the mild CLIP-T decline with increasing frame index and shows that, under multi-subject interaction, expressing every detailed verb perfectly becomes more difficult as the narrative horizon grows longer.

11. User Study

11.1. Protocol

We deploy a web-based survey interface, shown in Fig. 19. Each trial displays a grid of images generated from one long

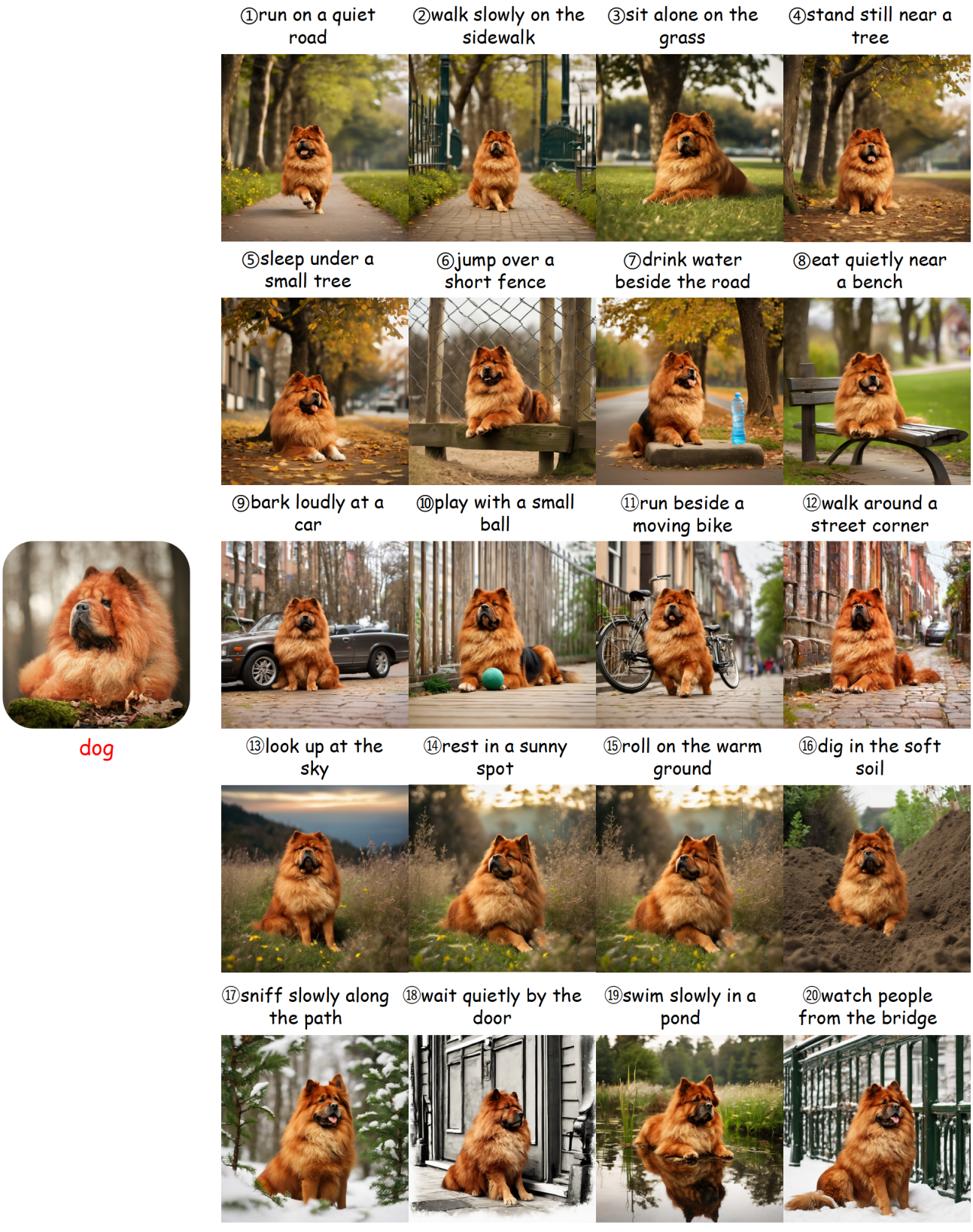


Figure 17. Long-frame narrative effect experiment on single-subject tasks (a 20-frame example).

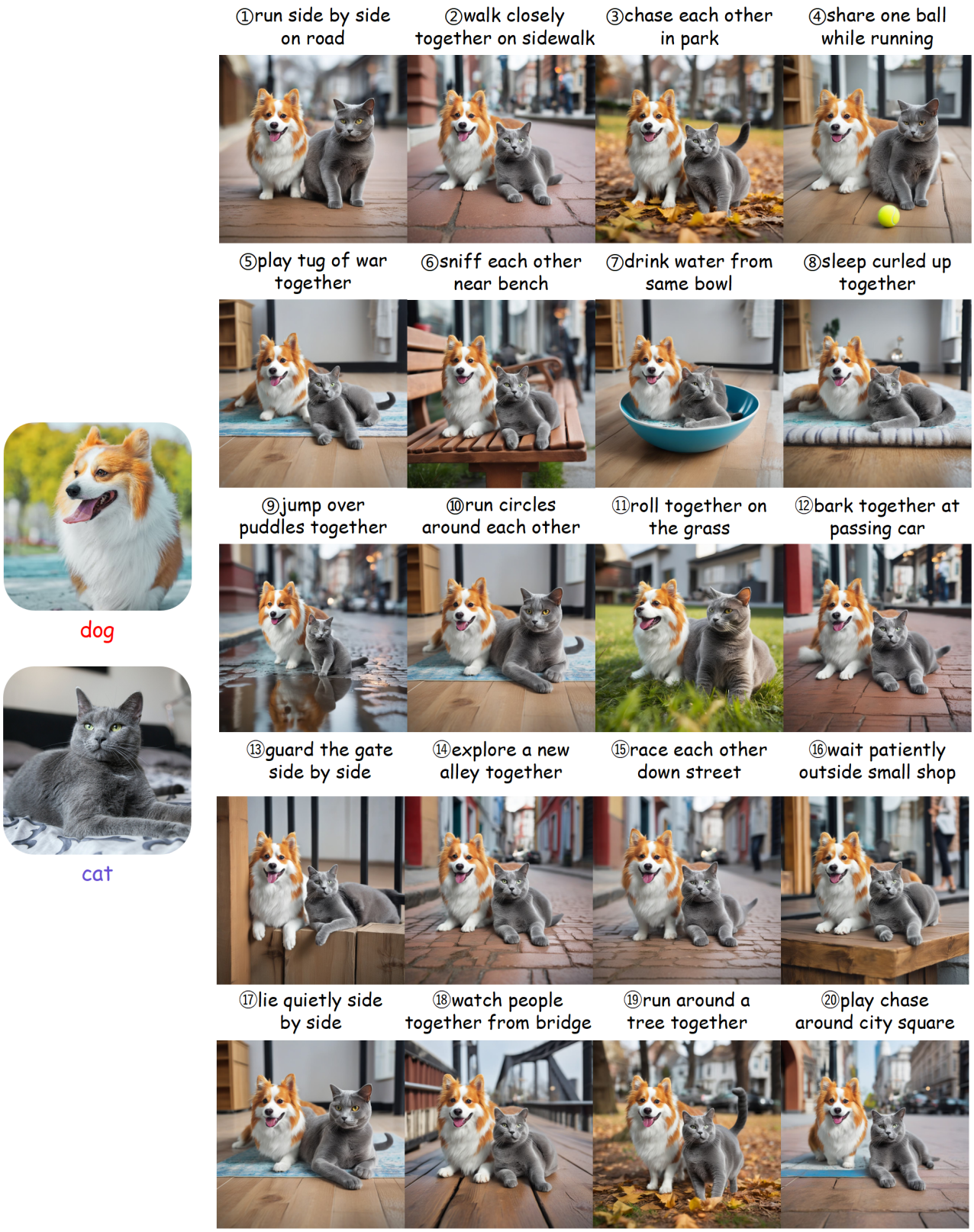


Figure 18. Long-frame narrative effect experiment on multi-subject tasks (a 20-frame example).

prompt by several anonymous methods: StoryTailor and the baselines in Sec. 4.2–4.3. Single-subject trials use twelve dog-action scenes indexed L1–L12. Multi-subject trials use sixteen cat–dog interaction scenes indexed R1–R16, such as hugging, dancing, and hand-over. Each column corresponds to one model; column order and prompt order are shuffled for every participant, and model names never appear, giving a double-blind setting.

Participants answer four questions for each grid.

Q1 asks how natural and internally consistent the full set looks. Ratings follow a five-point Likert scale that ranges from "Strongly disagree" to "Strongly agree".

Q2 asks whether multi-subject interactions look believable, with emphasis on poses, occlusion, and short-story coherence.

Q3 asks participants to pick a favourite single-subject image ID from L1–L12 and give a short reason, for instance smooth motion, stable lighting, or good alignment with the prompt.

Q4 asks for a favourite multi-subject image ID from R1–R16 and a short explanation, such as natural pose exchange, sensible occlusion, or vivid storytelling.

We recruit 100 volunteers, mainly graduate students and researchers in vision and graphics. Every participant completes all single- and multi-subject trials.

11.2. Analysis and Findings.

Single-subject ratings. Fig. 20(a) shows the Likert scores for single-subject prompts. Our method reaches a net agreement of about 72/%, clearly ahead of all baselines. Qwen-Edit is the strongest competitor with roughly 62% net agreement. FluxKontext follows at about 49%, while MS-Diffusion, IP-Adapter, and DreamBooth obtain around 36%, 23%, and 10%, respectively. The distribution of bars also shifts to the right for our model: most responses fall into "agree" or "strongly agree", and the fraction of negative ratings is the smallest among all methods. These trends suggest that StoryTailor preserves identity well and expresses the action cues in the prompt more faithfully, even when only one subject is present.

Multi-subject ratings. For multi-subject scenes in Fig. 20(b), the gap widens. Our method again achieves about 72% net agreement, while Qwen-Edit reaches 64%. Nano Banana and FluxKontext obtain roughly 53% and 42%, and MS-Diffusion and Lambda-Eclipse lag behind with about 31% and 20%. Reviewers report that our results have cleaner backgrounds, fewer collisions between subjects, and more coherent pose progressions across the sequence. Baselines more often receive neutral or negative scores due to background drag, attribute spill, or awkward contact between subjects.

Pairwise preference. Pairwise comparisons in Fig. 21 confirm these findings. On single-subject prompts, our method attains an overall preference score of about 70%, while Qwen-Edit reaches 62%, FluxKontext 55%, and MS-Diffusion, IP-Adapter, and DreamBooth stay between 40% and 47%. On multi-subject prompts, our method still leads with about 68% preference, followed by Qwen-Edit with 60%, Nano Banana with 56%, FluxKontext with 48%, and MS-Diffusion and Lambda-Eclipse with 42% and 38%. In qualitative comments, participants frequently describe our images as "more natural interactions" and "less confusing when subjects are close to each other".

StoryTailor attains the highest mean opinion score for both overall naturalness and interaction believability, with the clearest gains in multi-subject scenes. Participants often describe our outputs as having cleaner backgrounds, more natural pose exchanges, and smoother frame-to-frame progression. Baselines more frequently show background drag, attribute leakage, or awkward contact between subjects. In pairwise preference, StoryTailor wins the majority of comparisons against MS-Diffusion, FluxKontext, Qwen-Edit, NanoBanana and other strong baselines, especially in crowded or heavily occluded scenes. These subjective trends match the quantitative metrics: GCA and SFC reduce background carry-over and close-range confusion, while AB-SVR sharpens verb and interaction cues, leading to visual narratives that human viewers consistently prefer.

12. Limitations

Although StoryTailor improves action expression and multi-subject coherence under a 24 GB budget, several limitations remain. First, our design and evaluation are tied to SDXL with MS-Diffusion-style multi-subject conditioning and CLIP-family metrics, so generalization to other backbones, styles, and prompt distributions is not fully verified. Second, we mainly study short to medium narratives with 2–20 frames; as the sequence length grows, text embedding noise accumulates and CLIP-T gradually degrades, indicating that long-range story structure and very long narratives are not yet fully handled. Third, the current GCA and AB-SVR hyperparameters are hand-tuned for a limited set of subjects and actions, and performance can drop under extreme poses, heavy occlusions, dense crowds, or very loose bounding boxes. Fourth, SFC assumes moderate background continuity and may either over-smooth or under-propagate context in scenes with abrupt layout changes, fast camera motion, or strong lighting transitions. Finally, we rely on a small-scale user study and automatic metrics that only approximate human judgment, so a more systematic perceptual evaluation and task-specific benchmarks are left for future work.

13. Social Impacts

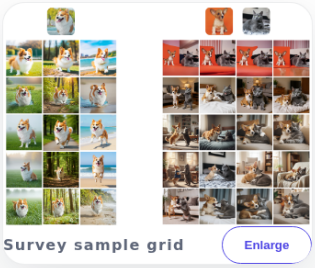
By making multi-subject, action-rich visual narratives feasible on a single commodity GPU, StoryTailor can support positive applications in education, digital heritage, pre-visualization, assistive storytelling, and personal creativity, especially for users without access to large GPU clusters. At the same time, the system operates on personal reference images and identity-consistent generation, which raises clear risks around privacy, consent, and potential misuse for deepfakes, harassment, or misleading narrative content. Our pipeline does not alter or weaken the safety mechanisms of the underlying diffusion backbone, but it may inherit and amplify dataset biases, stereotypes, or unequal representation patterns present in the pretrained models. Generated narratives could unintentionally reinforce cultural or gender stereotypes in how actions, roles, and relationships are depicted, particularly in multi-subject scenes. We therefore recommend deploying StoryTailor together with permission management for reference images, content filters, and human review in sensitive domains, and encourage future work on fairness-aware training data, controllable safety constraints, and tools that help users detect and mitigate harmful or deceptive generations.

QUESTIONNAIRE

Subjective rating form

This mock survey demonstrates how a static site could gather qualitative feedback. Any choice updates the tracker.

0 / 4 answered



Survey sample grid Enlarge

Answering tips

Single-subject IDs use L1-L12. Multi-subject IDs use R1-R16. Draft short notes before submitting.

After choosing an ID, jot 2–3 keywords explaining why it stood out so later discussions stay focused.

Q1 · Do the samples feel natural overall?

Score the collective realism and detail consistency across every ID.

- ○ Strongly disagree
- ○ Disagree
- ○ Neutral
- ○ Agree
- ○ Strongly agree

Q2 · Are the interactions between subjects believable?

Focus on the multi-subject scenes and judge the poses, occlusion, and storytelling.

- ○ Strongly disagree
- ○ Disagree
- ○ Neutral
- ○ Agree
- ○ Strongly agree

Q3 · Pick your favorite single-subject sample (L column)

Select the corgi render (L1-L12) that best matches the prompt and explain why.

Select an ID

Reason / notes: e.g., smooth motion, consistent lighting, prompt-aligned framing

Q4 · Pick your favorite multi-subject sample (R column)

Choose the cat-dog interaction (R1-R16) you prefer and highlight why it works.

Select an ID

Reason / notes: e.g., natural pose exchange, sensible occlusion, vivid story

Submit (demo) Clear

Figure 19. The web questionaire of StoryTailor about visual narrative tasks for users.

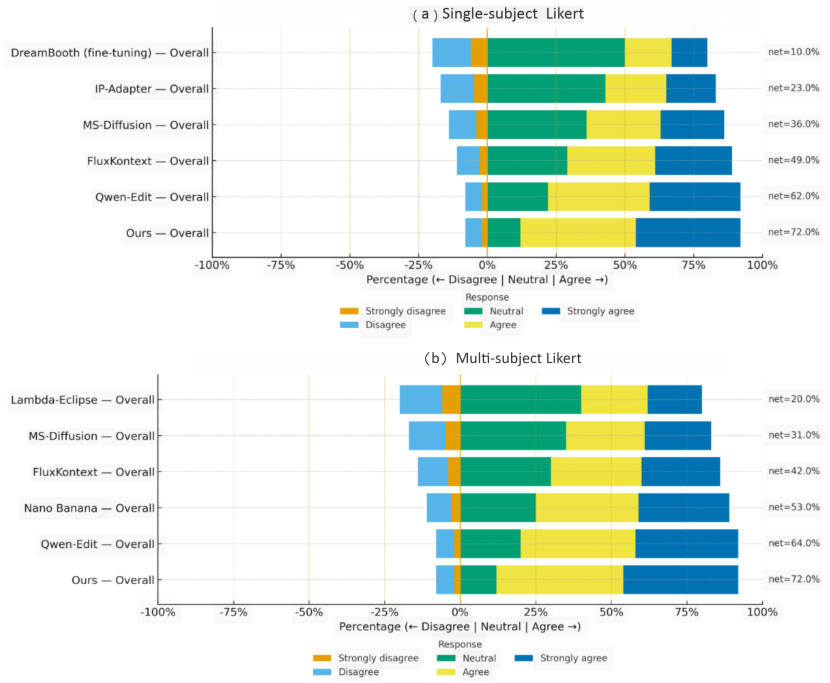


Figure 20. The Likert scores for single-subject(a) and multi-subject(b) tasks for user study

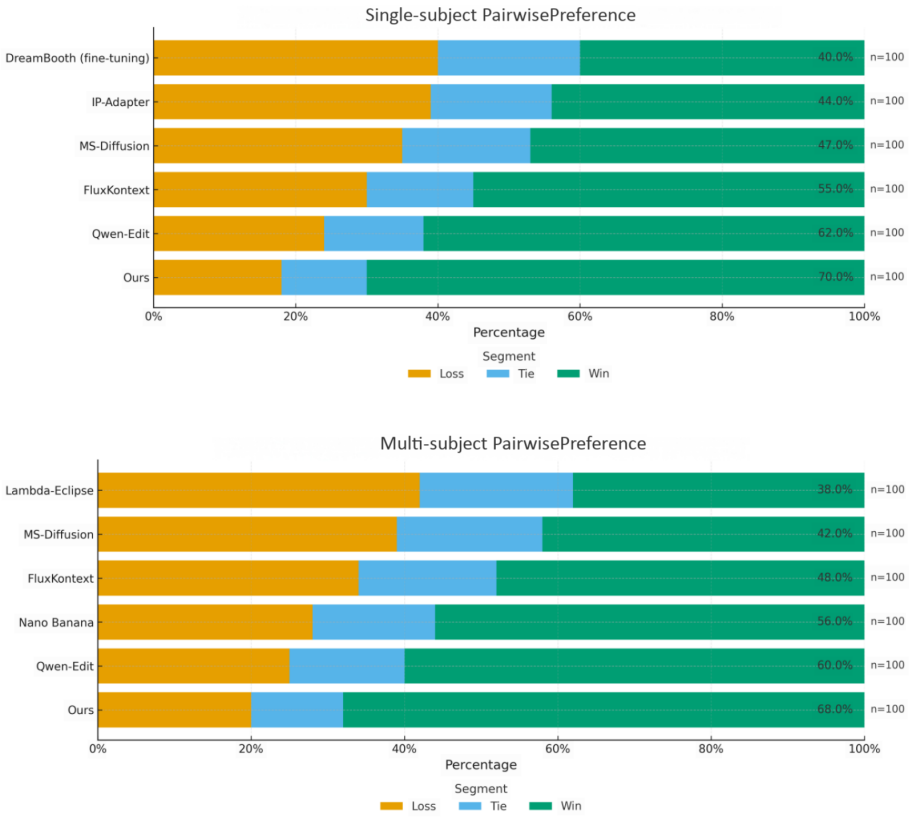


Figure 21. The pairwise preference for single-subject(a) and multi-subject(b) tasks for user study