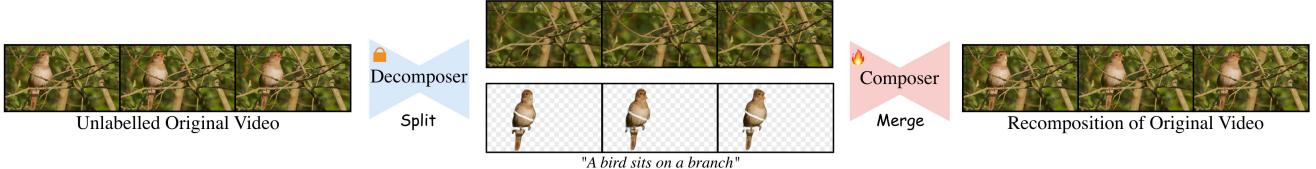


# Layer-Aware Video Composition via Split-then-Merge

Ozgur Kara<sup>1†</sup> Yujia Chen<sup>2</sup> Ming-Hsuan Yang<sup>2</sup>  
 James M. Rehg<sup>1</sup> Wen-Sheng Chu<sup>2‡</sup> Du Tran<sup>2‡</sup>

<sup>1</sup>University of Illinois Urbana-Champaign <sup>2</sup>Google

(a) Training: Split-then-Merge



(b) Inference: Generative Video Composition

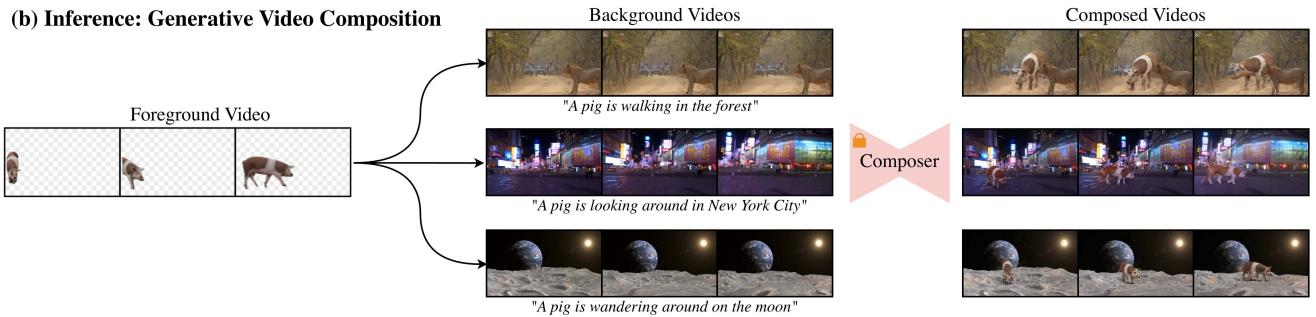


Figure 1. **Video Composition via Split-then-Merge.** (a) **Training:** The Decomposer splits an unlabeled video into foreground and background layers and generates a caption, while the Composer learns to merge them for reconstruction. (b) **Inference:** The Composer integrates a foreground video into novel background videos, and ensures affordance-aware placement (e.g., a pig on a forest road, NYC walkway, or lunar surface) with realistic harmonization (motion, lighting, shadows). Best viewed in color.

## Abstract

We present Split-then-Merge (*StM*), a novel framework designed to enhance control in generative video composition and address its data scarcity problem. Unlike conventional methods relying on annotated datasets or handcrafted rules, *StM* splits a large corpus of unlabeled videos into dynamic foreground and background layers, then self-composes them to learn how dynamic subjects interact with diverse scenes. This process enables the model to learn the complex compositional dynamics required for realistic video generation. *StM* introduces a novel transformation-aware training pipeline that utilizes a multi-layer fusion and augmentation to achieve affordance-aware composition, alongside an identity-preservation loss that maintains foreground fidelity during blending. Experiments show *StM* outperforms SOTA methods in both quantitative benchmarks and in humans/VLLM-based qualitative evaluations. More details are available at our [project page](#).

## 1. Introduction

Recent advances in diffusion models [3–7] have significantly improved the realism of synthesized videos. However, practical control over video generation remains largely restricted to text [8–13] and image conditioning [14–21], which remain too coarse for precise compositional guidance. Structured signals such as pose [22–25] or motion [26–28] provide stronger control, yet most methods are tuned for animating static images rather than flexible generative synthesis. Professional content creation, which fundamentally relies on layer-based compositing workflows [29, 30], demands more powerful compositional control. This motivates the task of *generative video composition*: synthesizing a single, coherent video by integrating a dynamic foreground video layer with a separate background video layer through the lens of modern generative AI. Unlike classical video composition [31], which primarily focuses on optimizing matting and blending, generative video composition allows for subtle adjustments to the foreground object—such as motion, shadowing, or pose—to harmonize it with the new environment’s lighting and camera dynamics. Figure 1(b) illustrates this compositional flexibility:

<sup>†</sup>Work done during an internship at Google

<sup>‡</sup>Joint last authors

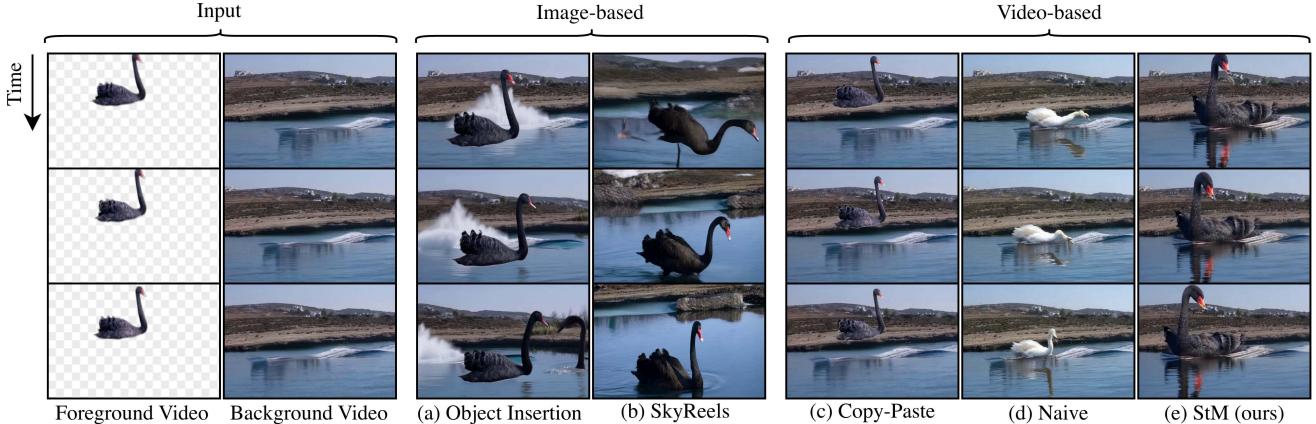


Figure 2. **Video Composition.** Given input foreground and background videos, image-based methods (a)–(b) use only the first frame, while (c)–(e) take full video inputs. (a) Object insertion [1] followed by Image-to-Video (I2V) and (b) end-to-end I2V composition SkyReels [2] fails to retain motion due to lack of video access. (c) Manual copy-paste preserves motion but violates affordance (swan placed on ground). (d) Naive generative composition yields appearance and motion drift (*e.g.*, black swan turns white). (e) Our method preserves identity and motion, and achieves affordance-aware placement with realistic blending (swan placed in water with wave and shadows).

the pig’s motion and appearance are adapted, and appropriate shadows are synthesized to match different backgrounds with varying lighting conditions and camera dynamics.

One may ask: “*How challenging is this task?*” or “*Can it be solved by simply adopting existing approaches?*”. A straightforward adaptation is leveraging an image-based object insertion and an image-to-video (I2V) approach. Specifically, the first frames of the foreground and background videos are fed into a state-of-the-arts (SOTA) object insertion method [1] to obtain a composed image, which is then inputted to an I2V model [10] to produce a video. As shown in Figure 2(a), this approach entirely discards the rich temporal information from the foreground video, forcing the I2V model to hallucinate motion from a text prompt alone. More recent approaches like SkyReels [32], which can directly compose a video from foreground and background images, suffers from the same limitation: missing motion information from the input, as shown in Figure 2(b).

“*Can the task be easily solved by simply switching to video inputs?*” Unfortunately, the answer is no. A simple copy-and-paste approach fails to be affordance-aware, often placing objects in incorrect locations. Figure 2(c) presents an example of this baseline where a swan is placed on the ground instead of in the water. A further attempt is to fine-tune a SOTA video generator [10] for this task, assuming sufficient data and annotations are provided. We found that a model naively trained on video composition still suffers from affordance issues as it learns easy shortcuts. Furthermore, this naive adaptation struggles with motion and appearance preservation. As shown in Figure 2(d), the swan’s appearance incorrectly changes to white, failing the desired effect of maintaining consistency with the original foreground. These experiments confirm that video composition is non-trivial and requires specialized solutions

beyond adopting existing off-the-shelf methods.

In this paper, we propose Split-then-Merge (StM), a novel, generic data-driven framework for video composition that requires *zero* manual annotation (Figure 1). Our key insight is to decompose unlabeled videos into layers and train a model to reconstruct the original video from them. This self-composition approach scales to any large unlabeled video dataset (*e.g.*, Panda-70M [33]). Unlike application-specific editing methods (*e.g.*, personalization [34], object-swapping [35], motion control [26–28, 36]), StM is purely data-driven and generalizes across diverse object types given a reliable decomposer. To address key composition challenges, we introduce two technical novelties: *a transformation-aware training pipeline* and *an identity-preservation loss*. The former stimulates affordance awareness by preventing trivial shortcut learning, while the latter balances foreground identity preservation with harmonious integration into the new background.

We establish comprehensive quantitative metrics for appearance and motion consistency, and introduce Vision-Language Large Models (VLLMs) as automated judges to augment user studies for qualitative evaluation. Our main contributions are:

- We propose StM, a scalable video composition framework that eliminates the need for manual annotations or additional conditions.
- To address the scarcity of training data, we release, StM-50K, the first multi-layer video dataset generated via a novel construction pipeline.
- Our comprehensive evaluation, including strong baselines and VLLM-based judges, demonstrates that StM significantly outperforms alternative methods.

## 2. Related Work

**Controllable Video Generation** The success of diffusion models in image synthesis [3–5] spurred rapid advancements in video generation, shifting the focus from quality to user control. Common modalities like text-guidance [8–13, 37] offer high-level semantic direction but lack precise control over subject identity or complex motion. While image-guided (I2V) models [14–21] enhance identity control via reference images, their static conditioning cannot convey dynamics. Likewise, structural signals such as human pose [22–25] or spatial conditions [17, 38–42] provide fine-grained control but are limited to image-level character animation rather than holistic synthesis. These approaches are thus constrained by their image-centric nature, controlling only isolated attributes. In contrast, our work addresses the more complex challenge of affordance-aware composition of the identity and motion from a complete foreground video with the dynamics of a background video.

**Image and Video Composition** Traditional composition techniques rely on pixel-level blending via color transfer [43, 44], harmonization [45], or gradient-domain pasting [46, 46] for realism [47, 48]. Video extensions introduced motion awareness [31, 49], but lacked generative capabilities. Generatively harmonizing two dynamic video layers remains largely unaddressed. Common baselines adapt image-level object insertion (*e.g.*, PbE [50], AnyDoor [51], Qwen-Edit [1]) and animate the result with an I2V model. These methods, including SkyReels [32], are fundamentally limited as conditioning on static images discards the foreground’s original motion. Other methods are distinct: AnyV2V [52] propagates single-image edits, while VideoAnyDoor [36] inserts a static image into a video. The most related work, LayerFlow [2], generates all layers from text, creating them from scratch. In contrast, StM is designed to compose two existing video layers, focusing on faithfully preserving their original identity and motion.

## 3. Video Composition via Split-then-Merge

### 3.1. Generative Video Composition

Generative video composition aims to synthesize a coherent new video by integrating a subject—including its appearance and motion—from one source video into the dynamic scene of another. Formally, the inputs to our model include: (i) a foreground video,  $V_{fg} \in [0, 255]^{T \times H \times W}$  (for simplicity we omit the color channel C), containing a primary subject (*e.g.*, a person, animal, or object), along with its corresponding binary segmentation mask  $M_{fg} \in \{0, 1\}^{T \times H \times W}$ ; (ii) a background video,  $V_{bg}$ , providing the scene context (*e.g.*, a beach or park); and (iii) a text prompt,  $\mathcal{T}$ , describing the desired final scene. The goal is to generate a video  $V_{pred}$  that seamlessly integrate the subject from  $V_{fg}$  into  $V_{bg}$  while maintaining visual and motion consistency.

This task presents two primary challenges. First, *layer consistency* requires preserving the distinct characteristics of both input layers. This involves maintaining the visual appearance and motion of the foreground subject (*e.g.*, a person’s gait) and the background scene (*e.g.*, scene specific and camera motion). Harmonizing these elements to avoid a trivial “pasted-on” look is essential for a coherent final video. Second, *affordance-awareness* requires a semantic understanding of the scene to ensure physically plausible placement and interaction—for instance, ensuring a car appears on a road rather than in the sky.

### 3.2. Split-then-Merge Framework

Split-then-Merge (StM) is a generic, data-driven framework for generative video composition built on the principle of *self-composition*. The core idea is to leverage off-the-shelf models to *split* an unlabeled video  $V_{org}$  into layers, and train a model to *merge* them back for reconstruction, as illustrated in Figure 1(a). Without human intervention, StM Decomposer can automatically *split* a vast corpus of unlabeled videos into their constituent layers: a foreground subject and a background scene, and generates a corresponding text caption. This process effectively transforms any unlabeled video into a training sample and enables the creation of a large-scale, multi-layer video dataset. During training, a diffusion-based generative model (*i.e.*, the Composer) is then trained to perform the inverse operation: learning to *merge* these layers back to the original, coherent video.

This approach stands in stark contrast to conventional methods in generative video synthesis and computational photography, which often rely on specialized algorithms, task-specific heuristics, or significant domain knowledge [2, 9, 52–55]. Such methods are frequently tailored to narrow domains (*e.g.*, face swapping [55–57]) that lack scalability and generality to handle diverse, in-the-wild content. By formulating the problem as inverting a decomposition process, our framework avoids these limitations, enabling a unified model that learns the fundamental principles of realistic video composition directly from data.

### 3.3. Decomposer: Splitting Videos into Layers

A core challenge for generative video composition is the absence of large-scale, multi-layered video datasets for training. To overcome this data scarcity problem, we develop an automated pipeline that decomposes standard video collections into four constituent layers: a text caption, a foreground video layer, a corresponding foreground mask, and a background video layer. StM Decomposer (presented in Figure 3), first employs a pre-trained video-language model [58] to generate a descriptive caption for a given video. Subsequently, an automatic motion segmentation model (*e.g.*, Segment-Any-Motion [59]) identifies and masks the primary moving subject within the scene. These

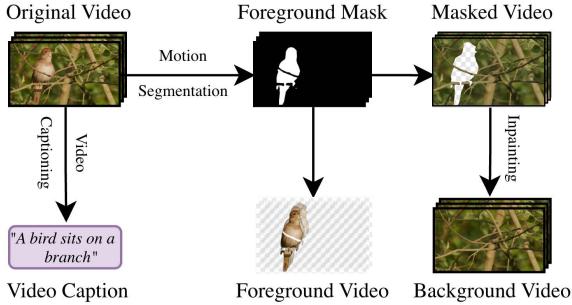


Figure 3. **StM Decomposer.** The StM Decomposer integrates off-the-shelf models to split unlabeled videos. First, motion segmentation generates a foreground mask, which is used to extract the foreground layer. An inpainting model then fills the “holes” in the masked background video. Finally, a video captioning model generates a descriptive text caption for the original video.

generated masks are then used to extract the foreground layer. To create a clean background layer, the masked region is removed from the original video, and a state-of-the-art video inpainting model [60] plausibly fills the resulting void, reconstructing a coherent background.

### 3.4. Composer: Learning to Merge

The StM Composer builds upon a latent diffusion transformer [10] originally developed for text-to-video generation. This framework comprises a pre-trained space-time VAE encoder-decoder ( $E_v$ ,  $D_v$ ), a text encoder ( $E_T$ ), and a Diffusion Transformer (DiT) backbone  $f(x_t; \theta)$ , where  $x_t$  is the latent input to DiT,  $\theta$  are DiT parameters, and  $t$  the diffusion timestep. The standard training objective is to recover the clean ground-truth latent from a noisy latent, conditioned on a text embedding. We adapt this architecture for video composition by introducing a transformation-aware training pipeline with three key components: multi-layer conditional fusion, transformation-aware augmentation, and a novel identity-preservation loss.

**Multi-layer Conditional Fusion** Figure 4 illustrates our proposed StM Composer architecture. First, the foreground video  $V_{fg}$  undergoes an augmentation operation to produce an augmented foreground  $\tilde{V}_{fg}$ . The original ground-truth, background, and augmented foreground videos are inputted into the encoder  $E_v$  to extract different latents:  $\mathbf{z}_{org}$ ,  $\mathbf{z}_{bg}$ , and  $\tilde{\mathbf{z}}_{fg}$  where  $\mathbf{z}_i = E_v(V_i)$  for all  $V_i \in \{V_{org}, V_{bg}, \tilde{V}_{fg}\}$ . The ground-truth latent  $\mathbf{z}_{org}$  is added with noise  $\epsilon_t$  and then fused other visual latents via a channel-wise concatenation and an MLP projection to produce a visual representation:

$$\mathbf{z}_{vision,t} = \text{MLP}(\text{Concat}_{\text{channel}}(\mathbf{z}_{org,t} + \epsilon_t, \tilde{\mathbf{z}}_{fg}, \mathbf{z}_{bg})), \quad (1)$$

where  $t$  is a specific time step during diffusion process. This *channel-level conditioning* is a critical design choice. Unlike methods that inject conditioning information as separate input tokens or through cross-attention mechanisms,

channel-wise fusion provides a dense, spatio-temporally-aligned guidance signal. It forces the model to consider the foreground content, background scene, and the noisy input simultaneously at every single spatio-temporal location within the latent space. This spatial alignment is crucial for recomposition, enabling precise, localized decisions, and realistic composition of the layers. Concurrently, the text prompt  $\mathcal{T}$  is encoded into a sequence of text embeddings using the text encoder,  $\mathbf{z}_{text} = E_T(\mathcal{T})$ . The final input to the transformer, which realizes the full condition, is denoted as  $\mathbf{x}_t$ . It is formed by reshaping the visual representation into a sequence of tokens and concatenating it with the text embeddings along the token dimension:

$$\mathbf{x}_t = \text{Concat}_{\text{token}}(\text{Tokens}(\mathbf{z}_{vision,t}), \mathbf{z}_{text}). \quad (2)$$

This combined representation serves as the full input to our DiT model, allowing it to leverage both fine-grained visual cues and high-level semantic guidance to perform the denoising task effectively, e.g., predicting the composed latent  $\mathbf{z}_t = f(\mathbf{x}_t; \theta)$  that reverts the process to match the original ground-truth latent  $\mathbf{z}_0 \triangleq \mathbf{z}_{org}$ .

**Transformation-Aware Augmentation** To prevent “copy-paste” shortcuts and encourage genuine compositional understanding, we introduce a targeted data augmentation strategy. The core idea is to make the recomposition task more challenging during training. By applying random transformations exclusively to the conditioning foreground layer,  $V_{fg}$ , we force the model to do more than just reconstruct; it must learn to *invert* these transformations to place the subject correctly and plausibly within the background scene. Specifically, our augmentations include spatial transformations like random horizontal flipping and random cropping and resizing, as well as photometric transformations such as color jittering. By systematically altering the subject’s orientation, scale, position, and color properties in the input condition, we create a more challenging training objective. This process prevents the model from memorizing fixed arrangements and is crucial for developing genuine *affordance-awareness* and ensuring *color harmony*. It teaches the model the underlying principles of how objects should be placed and colored to fit a scene, rather than just replicating a pattern.

**Identity-Preservation Loss** The standard loss from video latent diffusion transformer is:

$$\mathcal{L}_{\text{recon}} = \mathbb{E}_{\mathbf{z}_0,t} [\|\mathbf{z}_0 - f(\mathbf{x}_t, t; \theta)\|^2]. \quad (3)$$

We omit the constant weight  $w_t$  in Eq. (3) for simplicity. Standard reconstruction loss is effective for general text-to-video generation by treating all latent locations equally. In other words, because the encoder  $E_v$  preserves the relative spatiotemporal locality of features, applying a uniform

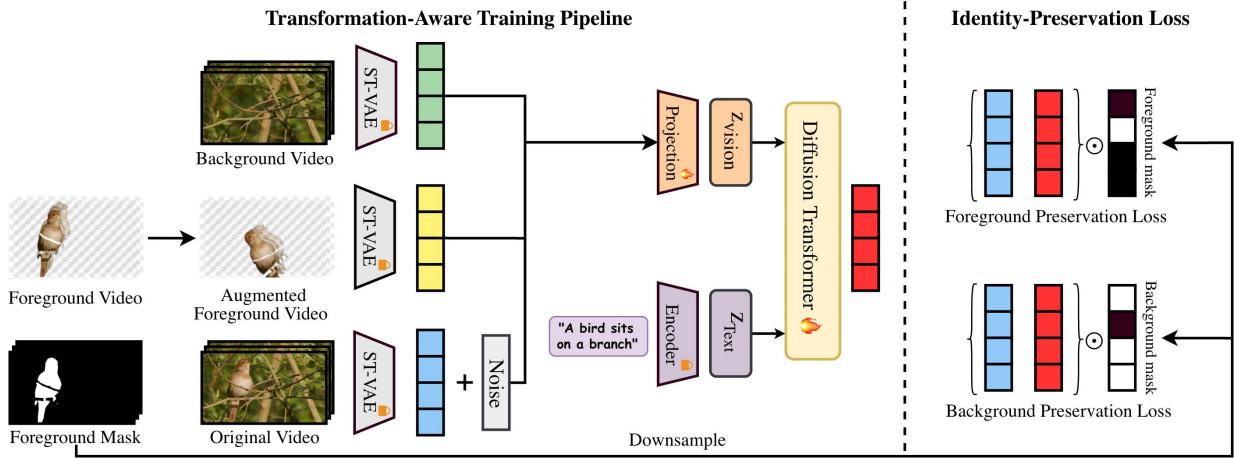


Figure 4. **StM Composer Training.** The Composer is trained to reconstruct a ground-truth video latent from foreground, background, and text inputs. First, the foreground video is augmented, and all video inputs (augmented foreground, background, ground truth) are encoded into latents by a frozen Space-Time (ST) VAE. The text prompt is encoded as  $Z_{text}$ . A noisy ground-truth latent (blue) is fused with background (green) and augmented foreground (yellow) latents via a projection layer to produce the visual representation  $Z_{vision}$ . A Diffusion Transformer then processes  $Z_{vision}$  and  $Z_{text}$  to predict a composed latent (red). The identity-preservation loss comprises two weighted sub-losses comparing the prediction (red) against the ground truth (blue) using foreground- and background-aware masking.

L2 loss implicitly weighs all spatial regions equally. However, video composition requires a delicate balance between foreground identity preservation and harmonious synthesis. Naively optimizing for harmony can degrade the foreground object’s identity (as shown in Figure 2(d)), while strictly enforcing identity preservation often results in unnatural, disjointed composition lacking visual harmony.

We propose an identity-preservation loss that balances foreground object identity retention and overall compositional harmony. Using the training-time foreground mask  $M$ , we decouple the latent space into foreground and background regions with separate weights, thereby decomposing the reconstruction loss into two sub-losses:

$$\begin{aligned} \mathcal{L}_{fg} &= \mathbb{E}_{\mathbf{z}_0, t} \left[ \frac{\sum((\mathbf{z}_0 - \mathbf{f}(\mathbf{x}_t, t; \theta))^2 \odot M)}{\sum M} \right], \\ \mathcal{L}_{bg} &= \mathbb{E}_{\mathbf{z}_0, t} \left[ \frac{\sum((\mathbf{z}_0 - \mathbf{f}(\mathbf{x}_t, t; \theta))^2 \odot (1 - M))}{\sum(1 - M)} \right]. \end{aligned} \quad (4)$$

We note that both foreground and background sub-losses are normalized by their respective pixel area (*i.e.*, the size of the foreground mask  $M$  and the background area). The final identity-preservation loss is then computed as a weighted sum of these two normalized terms:  $\mathcal{L}_{final} = \alpha \mathcal{L}_{fg} + (1 - \alpha) \mathcal{L}_{bg}$ . This design provides direct control over the trade-off between foreground identity preservation and overall composition quality, and it inherently mitigates the sensitivity to varying foreground object sizes observed during training.

**Inference.** Inference proceeds similarly to training, with three key exceptions: (i) the ground-truth latent is replaced by a noise latent  $\epsilon$ ; (ii) augmentation is omitted for the foreground video; and (iii) the space-time decoder ( $D_v$ ) is applied to map the predicted latent back to the pixel space.

## 4. Experiments

### 4.1. Implementation Details

**Decomposer** Our Decomposer pipeline (Section 3.3) processes unlabeled videos using off-the-shelf models. It utilizes InternVL [61] for video captioning, Segment-Any-Motion [59] for motion segmentation to extract foreground masks, and MiniMax Remover [60] for video inpainting to generate clean background layers. Our training dataset, StM-50K, comprises 50K video clips pre-processed by this Decomposer. To ensure robust performance, we processed several sources. For unannotated collections like Panda-70M [33] and Animal Kingdom [62], we apply our full pipeline. For annotated datasets such as Youtube-VOS [63] and LVOS [64], we adapt our pipeline to leverage the provided ground-truth foreground masks. The DAVIS [65] dataset is reserved exclusively for validation. For evaluation, we curate a dedicated test benchmark of 93 unique, unseen triplets (foreground video, background video, text prompt) from our held-out set. Each sample’s foreground and background are intentionally sourced from different videos, measuring the model’s ability to generate plausible interactions for unseen combinations.

**Composer** We adopt the CogVideoX-I2V model [10] as our base architecture, initializing it with pre-trained weights and fully fine-tuning for 20K iterations. Training is conducted on 16 NVIDIA H100 GPUs with a total batch size of 64, using bf16 mixed precision. We employ the AdamW optimizer [66] (with  $\beta_1 = 0.9$ ,  $\beta_2 = 0.95$ ,  $\beta_3 = 0.98$ ,  $\epsilon = 1e-8$ , weight decay  $1e-4$ , and max gradient norm 1.0) and a cosine learning rate scheduler (base LR  $5e-6$ , 1000 warm-up steps, single cycle). All video clips are processed

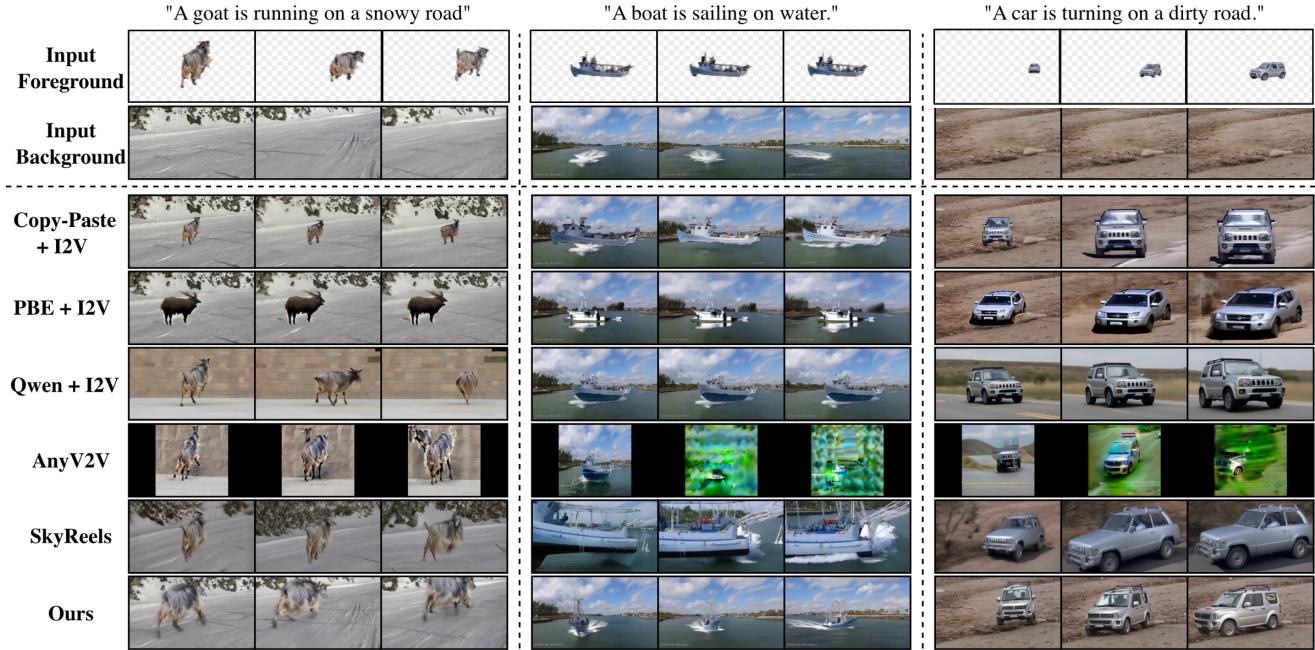


Figure 5. **Qualitative comparison.** Our method (StM) uniquely preserves complex dynamics and achieves affordance-aware harmony where baselines fail. (**Left**) StM alone maintains both the rapid background camera motion and realistic foreground running motion. (**Center**) StM demonstrates affordance by adapting the boat’s orientation and height of the waves. (**Right**) StM accurately preserves the car’s semantic action, road alignment, and lighting consistency, unlike alternative methods.

at a  $49 \times 480 \times 720$  resolution. The identity-preservation loss weight  $\alpha$  is set to 0.5. To mitigate shortcut learning, we apply a sequential data augmentation pipeline to the foreground: (1) *random horizontal flipping* ( $p = 0.7$ ), (2) *random resized cropping* ( $p=0.7$ ) with a scale range [0.5, 2.0] and ensures 90% of the original foreground is preserved, which simulates varied camera perspectives and ensures robustness to translation and rotation variances. (3) *color shifting* ( $p = 0.2$ ) exclusively to the masked foreground, randomly jittering brightness, contrast, saturation, and hue within a [0, 0.2] range to encourage color harmonization.

## 4.2. Evaluation Setup

**Baselines** We compare our approach against five baselines, grouped into *cascaded I2V methods* and *video-centric methods*. The three cascaded methods share a pipeline: they first compose the initial frames of the input videos at the image-level, and this single composite is then animated by our I2V base model. These baselines include: **Copy-Paste + I2V**, which naively scales the foreground bounding box to half the background size and pastes it into the center; **PBE + I2V**, which uses Paint-by-Example [50] to inpaint the foreground into a central bounding box; and **Qwen + I2V**, which employs SOTA Qwen-Edit [1] for image composition. Our second category includes **SkyReels** [32], an end-to-end framework that composes multiple static images and generates video, and **AnyV2V** [52], a motion-aware, tuning-free video editing framework. A key limitation of

Table 1. **Quantitative Metrics.** Our metrics are designed to evaluate the generative video composition across different focus areas.

Metric	Target	Measurement Method
(M1) Identity Preservation	FG / BG	ViCLIP [58] embedding similarity between the original input and segmented output layers (computed separately for FG and BG).
(M2) Semantic Action Alignment	FG	KL Divergence between action probability distributions (Video Swin [67]) of the input and output FG.
(M3) Background Motion Alignment	BG	Mean Squared Error (MSE) between Optical Flow fields [68] of the input and output BG.
(M4) Textual Alignment	Full	ViCLIP [58] embedding similarity between the final composite video and the guidance text prompt.

the first four baselines is their inability to preserve original foreground motion, as they all operate on static inputs.

**Metrics** Overall, we assess StM across 5 key criteria using both automated quantitative metrics and qualitative preference studies, including a user study and VLLM-as-a-Judge. For quantitative evaluation, we use [59] to decompose the generated video into separate Foreground (FG) and Background (BG) layers, which are then evaluated against the corresponding input layers using the pre-trained models and metrics provided in Table 1. While M1 (Identity Preservation) measures visual appearance consistency, the motion

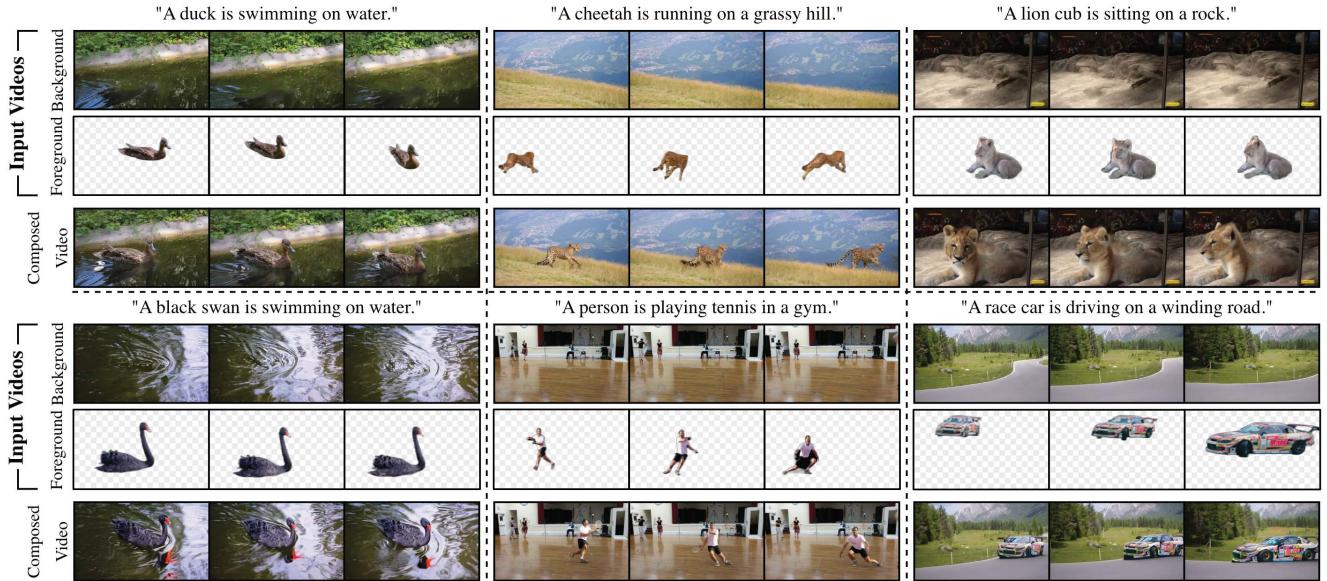


Figure 6. **Qualitative Results.** StM demonstrates robust motion preservation and affordance awareness. Characteristic actions are faithfully retained (*e.g.*, cheetah, tennis player). Subjects are plausibly integrated rather than “pasted,” as seen with the swan and duck adapted to water flow, and the lion and race car harmonized with appropriate scene lighting and placement.

consistency is evaluated at two levels: M2 (Semantic Action Alignment) assesses semantic-level motion (*i.e.*, action) for the FG, and M3 (Background Motion Alignment) assesses pixel-level motion (*i.e.*, camera & scene dynamics) for the BG. Finally, M4 (Textual Alignment) measures the coherence between the text prompt and the visual output.

For qualitative evaluation, we recruit 50 subjects on the Prolific platform [69] for the user study on randomly sampled 25 samples from our test dataset, and utilized Gemini 2.5 Pro (Public version) [70] as the VLLM judge (details in Appendix B) on the full test dataset. These criteria are: (i) ***Identity Preservation (FG & BG)***, which measures visual identity fidelity. (ii) ***Motion Alignment (FG & BG)***, assessing the fidelity of subject and camera/scene motion. (iii) ***Textual Alignment***, measuring prompt adherence via automated ViCLIP embedding similarity between the composite video and the input text. (iv) ***FG-BG Harmony***, assessing plausible interaction and affordance-awareness, evaluated exclusively in our qualitative studies. (v) ***Overall Quality***, a holistic judgment of realism, artifacts, and coherence, also evaluated exclusively qualitatively.

### 4.3. Analysis and Discussion

**Automated Quantitative Evaluation** As shown in Table 2, our automated evaluation confirms StM’s quantitative advantages. For ***Identity Preservation***, StM achieves the highest scores for both foreground (FG: 84.82) and background (BG: 92.88). This highlights how our full-video conditioning avoids the appearance drift that I2V baselines suffer when hallucinating dynamics from a single frame—a key issue for methods like PBE + I2V, which can struggle

Table 2. **Quantitative Evaluation.** Comparison with baseline methods. We report **M1** (Identity Preservation), **M2** (Semantic Action Alignment), **M3** (Background Motion Alignment), and **M4** (Textual Alignment). Metric directions are marked ( $\uparrow$ ,  $\downarrow$ ). M1 and M4 scores are multiplied by 100. **Bold** indicates the best performance, and underline is the second best.

Method	Identity Preserv.		Action / Motion		Textual Align. $\uparrow$
	FG $\uparrow$	BG $\uparrow$	FG $\downarrow$	BG $\downarrow$	
Copy-Paste + I2V	83.08	85.02	<u>1.61</u>	184.59	19.21
PBE + I2V	73.52	80.19	2.47	98.08	19.41
Qwen + I2V	82.02	72.38	1.71	<u>74.77</u>	24.20
SkyReels	80.24	75.24	1.75	279.23	<b>24.40</b>
AnyV2V	77.73	56.36	1.70	154.97	24.13
<b>StM (ours)</b>	<b>84.82</b>	<b>92.88</b>	<b>1.22</b>	<b>16.36</b>	19.81

with identity. This advantage is more pronounced in ***Action & Motion Alignment*** (lower is better). StM attains the best FG action alignment score (1.22) by a significant margin over baselines (1.61-2.47), as I2V methods must guess motion. The difference is stark for BG motion alignment: our score (16.36) is an order of magnitude better than all competitors (74.77-279.23), which fail to preserve original camera/scene dynamics and often produce static backgrounds. This quantitative failure reflects the qualitative issues in baselines like Copy-Paste + I2V, which lacks affordance-awareness, and Qwen + I2V, which can have content consistency issues. For ***Textual Alignment***, text-guided baselines like SkyReels (24.40) score higher than StM (19.81). This is an expected trade-off: StM prioritizes faithful preservation of the input videos’ motion and appearance as dominant

Table 3. **Pairwise Preference Study (Win Rates).** We present a qualitative study where evaluators and a VLLM performed a pairwise comparison, choosing between our method and a baseline. The percentages denote our method’s “win rate”—*i.e.*, how often it was preferred over the baseline. **For all metrics, higher values are better.**

Ours vs. Baseline	User Study								VLLM as a Judge (Gemini 2.5 Pro)									
	Identity Preserv.		Motion Align.		FG-BG Harmony	Overall Quality	Identity Preserv.		Motion Align.		FG-BG Harmony	Overall Quality	Identity Preserv.		Motion Align.		FG-BG Harmony	Overall Quality
	FG	BG	FG	BG			FG	BG	FG	BG			FG	BG	FG	BG		
Copy-Paste + I2V	72.73%	83.04%	76.65%	84.90%	86.53%	86.50%	57.30%	85.88%	65.17%	80.00%	91.95%	90.00%	57.30%	85.88%	65.17%	80.00%	91.95%	90.00%
PBE + I2V	87.77%	88.00%	84.76%	88.06%	88.54%	88.50%	87.78%	95.35%	84.27%	90.91%	90.80%	92.22%	87.78%	95.35%	84.27%	90.91%	90.80%	92.22%
Qwen + I2V	59.06%	73.56%	61.23%	70.44%	55.11%	55.10%	52.17%	90.36%	52.17%	84.81%	54.95%	46.73%	52.17%	90.36%	52.17%	84.81%	54.95%	46.73%
SkyReels	68.74%	82.38%	67.38%	79.59%	64.46%	64.50%	57.14%	94.44%	61.11%	89.87%	69.66%	65.93%	57.14%	94.44%	61.11%	89.87%	69.66%	65.93%
AnyV2V	84.45%	87.46%	88.80%	88.72%	89.62%	89.60%	96.10%	100.00%	98.72%	98.75%	95.00%	100.00%	96.10%	100.00%	98.72%	98.75%	95.00%	100.00%

Table 4. **Ablation Study.** Attributions of each component by removing it from our full model. Metrics are the same as Table 2.

Method	Identity Preserv.		Action / Motion		Textual Align. ↑
	FG ↑	BG ↑	FG ↓	BG ↓	
<b>StM</b>	84.82	<b>92.88</b>	1.22	<b>16.36</b>	<b>19.81</b>
w/o Aug.	83.15	89.20	0.70	16.61	16.56
w/o ID Loss	82.01	88.25	0.92	23.75	16.42
w/o Both	<b>90.39</b>	84.76	<b>0.77</b>	18.59	18.70

control signals, whereas text-guided baselines prioritize semantic adherence to the prompt over visual source fidelity.

**User Study / VLLM as a Judge** We present qualitative preference studies (Table 3) to support our automated quantitative evaluations (Table 2). Furthermore, these studies measure attributes like FG-BG Harmony and Overall Quality, which automated metrics cannot adequately capture. For Identity Preservation, the findings from both human and VLLM evaluators align with our automated metrics, confirming a consistent preference for StM. For Action & Motion Alignment (FG & BG), StM achieves exceptionally high win rates. This result is critical as it validates a core failure of the I2V-based baselines: by operating on static frames, they discard the input video’s dynamics, whereas StM, by conditioning on the full video layers, successfully preserves them. Similarly, StM is strongly preferred for FG-BG Harmony (often >85% win rates), confirming it learns plausible, affordance-aware interactions, which contrasts sharply with other methods that fail to harmonize the subject. The high win rates in Overall Quality directly reflect StM’s ability to solve these perceptual challenges. While Qwen + I2V is somewhat competitive in Overall Quality, likely due to its SOTA image editing base, it fails significantly across all motion and preservation metrics, suggesting its perceived visual quality does not account for these critical dynamic and identity failures.

**Ablation Study** Table 4 details our ablation study. The w/o Both baseline learns a detrimental “copy-and-paste” shortcut, memorizing the foreground’s position rather than learning composition. This yields misleadingly good FG scores (90.39 FG Identity, 0.77 FG Action) but poor com-

positional quality (BG Motion 18.59). Our transformation-aware augmentation breaks this shortcut: without *ID Loss*, it forces genuine composition, lowering the simplistic FG Identity score to 82.01 and slightly increasing FG Action error to 0.92. Conversely, removing *Augmentation* alone fails to break the shortcut, as FG Action remains at a near-perfect 0.70. The full StM model combines both components to achieve the best balance. While FG scores (84.82 FG Identity, 1.22 FG Action) naturally decrease compared to the shortcut-learning baseline, the full model achieves superior generalization metrics (92.88 BG Identity, 16.36 BG Motion), confirming that both components are essential for robust, affordance-aware composition.

**Qualitative Evaluation** Qualitative results (Figures 5, 6) confirm StM’s superior ability to generate affordance-aware, dynamically coherent compositions. Figure 5 compares methods, showing StM (Left) preserves complex elements like rapid camera movement with the goat’s running action, scene affordance by correctly orienting a boat and adapting its vertical position to waves (Center), and accurately preserving a car’s complex turn while aligning it with the road and adapting lighting (Right). Figure 6 further shows its robustness by faithfully retaining characteristic subject actions (cheetah, tennis player), while achieving deep integration by realistically adapting subjects (swan, duck, lion) to water flow, shadows, and scene lighting, preventing a “pasted” look.

## 5. Conclusion

We introduce StM, a unified framework for generative video composition designed to address data scarcity through a scalable “Split-then-Merge” paradigm. Specifically, by decomposing unlabeled videos into dynamic layers and self-composing them, StM learns to compose complex subject-scene without manual annotations. Powered by the novel StM-50K multi-layer dataset, our data-driven approach employs two critical components: an *identity-preservation loss* to maintain subject fidelity and *transformation-aware training* to capture realistic motion and affordances. Experiments show StM consistently outperforms baselines in preserving motion dynamics and ensuring harmonious com-

position compared to existing alternative methods. Future work will address limitations such as the trade-off between visual fidelity and textual alignment (see Appendix C).

## Acknowledgments

Portions of this research were supported in part by the Health Care Engineering Systems Center in the Grainger College of Engineering at UIUC, and the National Institutes of Health (NIH) under award P41EB028242.

## References

- [1] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuteng Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025. [2](#), [3](#), [6](#)
- [2] Sihui Ji, Hao Luo, Xi Chen, Yuanpeng Tu, Yiyang Wang, and Hengshuang Zhao. Layerflow: A unified model for layer-aware video generation. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–10, 2025. [2](#), [3](#)
- [3] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. [1](#), [3](#)
- [4] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.
- [5] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. [3](#)
- [6] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.
- [7] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. [1](#)
- [8] Ozgur Kara, Bariscan Kurtkaya, Hidir Yesiltepe, James M Rehg, and Pinar Yanardag. Rave: Randomized noise shuffling for fast and consistent video editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6507–6516, 2024. [1](#), [3](#)
- [9] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. In *The Twelfth International Conference on Learning Representations*, 2024. [3](#)
- [10] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazhen Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. In *The Thirteenth International Conference on Learning Representations*, 2025. [2](#), [4](#), [5](#)
- [11] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. In *The Eleventh International Conference on Learning Representations*, 2023.
- [12] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Hermann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, et al. Lumiere: A space-time diffusion model for video generation. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024.
- [13] Agrim Gupta, Lijun Yu, Kihyuk Sohn, Xiuye Gu, Meera Hahn, Fei-Fei Li, Irfan Essa, Lu Jiang, and José Lezama. Photorealistic video generation with diffusion models. In *European Conference on Computer Vision*, pages 393–411. Springer, 2024. [1](#), [3](#)
- [14] Xun Guo, Mingwu Zheng, Liang Hou, Yuan Gao, Yufan Deng, Pengfei Wan, Di Zhang, Yufan Liu, Weiming Hu, Zhengjun Zha, et al. I2v-adapter: A general image-to-video adapter for diffusion models. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–12, 2024. [1](#), [3](#)
- [15] Weiming Ren, Huan Yang, Ge Zhang, Cong Wei, Xinrun Du, Wenhao Huang, and Wenhui Chen. Consisti2v: Enhancing visual consistency for image-to-video generation. *Transactions on Machine Learning Research*, 2024.
- [16] Xiaoyu Shi, Zhaoyang Huang, Fu-Yun Wang, Weikang Bian, Dasong Li, Yi Zhang, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, et al. Motion-i2v: Consistent and controllable image-to-video generation with explicit motion modeling. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024.
- [17] Koichi Namekata, Sherwin Bahmani, Ziyi Wu, Yash Kant, Igor Gilitschenski, and David B. Lindell. SG-i2v: Self-guided trajectory control in image-to-video generation. In *The Thirteenth International Conference on Learning Representations*, 2025. [3](#)
- [18] Wenqi Ouyang, Yi Dong, Lei Yang, Jianlou Si, and Xingang Pan. I2vedit: First-frame-guided video editing via image-to-video diffusion models. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024.
- [19] Wanquan Feng, Jiawei Liu, Pengqi Tu, Tianhao Qi, Mingzhen Sun, Tianxiang Ma, Songtao Zhao, SiYu Zhou, and Qian HE. I2VControl-camera: Precise video camera control with adjustable motion strength. In *The Thirteenth International Conference on Learning Representations*, 2025.
- [20] Haoyu Zhao, Tianyi Lu, Jiaxi Gu, Xing Zhang, Qingping Zheng, Zuxuan Wu, Hang Xu, and Yu-Gang Jiang. Magdiff: Multi-alignment diffusion for high-fidelity video generation and editing. In *European Conference on Computer Vision*, pages 205–221. Springer, 2024.
- [21] Yuming Jiang, Tianxing Wu, Shuai Yang, Chenyang Si, Dahua Lin, Yu Qiao, Chen Change Loy, and Ziwei Liu.

- Videobooth: Diffusion-based video generation with image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6689–6700, 2024. 1, 3
- [22] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Siran Chen, Xiu Li, and Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4117–4125, 2024. 1, 3
- [23] Di Chang, Yichun Shi, Quankai Gao, Hongyi Xu, Jessica Fu, Guoxian Song, Qing Yan, Yizhe Zhu, Xiao Yang, and Mohammad Soleymani. Magicpose: Realistic human poses and facial expressions retargeting with identity-aware diffusion. In *Forty-first International Conference on Machine Learning*, 2024.
- [24] Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. Dreampose: Fashion video synthesis with stable diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22680–22690, 2023.
- [25] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8153–8163, 2024. 1, 3
- [26] Chong Mou, Mingdeng Cao, Xintao Wang, Zhaoyang Zhang, Ying Shan, and Jian Zhang. Revideo: Remake a video with motion and content control. *Advances in Neural Information Processing Systems*, 37:18481–18505, 2024. 1, 2
- [27] Daniel Geng, Charles Herrmann, Junhwa Hur, Forrester Cole, Serena Zhang, Tobias Pfaff, Tatiana Lopez-Guevara, Yusuf Aytar, Michael Rubinstein, Chen Sun, et al. Motion prompting: Controlling video generation with motion trajectories. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1–12, 2025.
- [28] 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. 1, 2
- [29] Ron Brinkmann. *The art and science of digital compositing: Techniques for visual effects, animation and motion graphics*. Morgan Kaufmann, 2008. 1
- [30] Steve Wright. *Digital compositing for film and video*. Routledge, 2013. 1
- [31] Tao Chen, Jun-Yan Zhu, A. Shamir, and Shi-Min Hu. Motion-aware gradient domain video composition. *Image Processing, IEEE Transactions on*, 22(7):2532–2544, 2013. 1, 3
- [32] Zhengcong Fei, Debang Li, Di Qiu, Jiahua Wang, Yikun Dou, Rui Wang, Jingtao Xu, Mingyuan Fan, Guibin Chen, Yang Li, et al. Skyreels-a2: Compose anything in video diffusion transformers. *arXiv preprint arXiv:2504.02436*, 2025. 2, 3, 6
- [33] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13320–13331, 2024. 2, 5, 13
- [34] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Yuwei Fang, Kwot Sin Lee, Ivan Skorokhodov, Kfir Aberman, Jun-Yan Zhu, Ming-Hsuan Yang, and Sergey Tulyakov. Multi-subject open-set personalization in video generation. 2025. 2
- [35] Yuchao Gu, Yipin Zhou, Bichen Wu, Licheng Yu, Jia-Wei Liu, Rui Zhao, Jay Zhangjie Wu, David Junhao Zhang, Mike Zheng Shou, and Kevin Tang. Videoswap: Customized video subject swapping with interactive semantic point correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7621–7630, June 2024. 2
- [36] Yuanpeng Tu, Hao Luo, Xi Chen, Sihui Ji, Xiang Bai, and Hengshuang Zhao. Videoanydoor: High-fidelity video object insertion with precise motion control. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Papers*, pages 1–11, 2025. 2, 3
- [37] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15954–15964, 2023. 3
- [38] Yuwei Guo, Ceyuan Yang, Anyi Rao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Sparsectrl: Adding sparse controls to text-to-video diffusion models. In *European Conference on Computer Vision*, pages 330–348. Springer, 2024. 3
- [39] Jiawei Wang, Yuchen Zhang, Jiaxin Zou, Yan Zeng, Guoqiang Wei, Liping Yuan, and Hang Li. Boximator: Generating rich and controllable motions for video synthesis. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 52274–52289. PMLR, 21–27 Jul 2024.
- [40] Jianzong Wu, Xiangtai Li, Yanhong Zeng, Jiangning Zhang, Qianyu Zhou, Yining Li, Yunhai Tong, and Kai Chen. Motionbooth: Motion-aware customized text-to-video generation. *Advances in Neural Information Processing Systems*, 37:34322–34348, 2024.
- [41] Weijia Wu, Zhuang Li, Yuchao Gu, Rui Zhao, Yefei He, David Junhao Zhang, Mike Zheng Shou, Yan Li, Tingting Gao, and Di Zhang. Draganything: Motion control for anything using entity representation. In *European Conference on Computer Vision*, pages 331–348. Springer, 2024.
- [42] Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, XIAOPENG ZHANG, Wangmeng Zuo, and Qi Tian. Controlvideo: Training-free controllable text-to-video generation. In *The Twelfth International Conference on Learning Representations*, 2024. 3

- [43] Erik Reinhard, Michael Adhikhmin, Bruce Gooch, and Peter Shirley. Color transfer between images. *IEEE Computer graphics and applications*, 21(5):34–41, 2001. 3
- [44] F. Pitie, A. C. Kokaram, and R. Dahyot. N-dimensional probability density function transfer and its application to color transfer. In *ICCV*, 2005. 3
- [45] Daniel Cohen-Or, Olga Sorkine, Ran Gal, Tommer Leyvand, and Ying-Qing Xu. Color harmonization. *ACM Transactions on Graphics*, 25(3):624–630, 2006. 3
- [46] Jiaya Jia, Jian Sun, Chi-Keung Tang, and Heung-Yeung Shum. Drag-and-drop pasting. *ACM Transactions on Graphics*, 25(3):631–637, 2006. 3
- [47] Jean-Francois Lalonde and Alexei A Efros. Using color compatibility for assessing image realism. In *ICCV*, 2007. 3
- [48] Su Xue, Aseem Agarwala, Julie Dorsey, and Holly Rushmeier. Understanding and improving the realism of image composites. *ACM Transactions on Graphics*, 31(4):84, 2012. 3
- [49] Jingye Wang, Bin Sheng, Ping Li, Yuxi Jin, and David Dagan Feng. Illumination-guided video composition via gradient consistency optimization. *Trans. Img. Proc.*, 28(10):5077–5090, October 2019. 3
- [50] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18381–18391, 2023. 3, 6
- [51] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6593–6602, 2024. 3
- [52] Max Ku, Cong Wei, Weiming Ren, Huan Yang, and Wenhui Chen. Anyv2v: A tuning-free framework for any video-to-video editing tasks. *Transactions on Machine Learning Research*, 2024. Reproducibility Certification. 3, 6
- [53] Boxiao Pan, Zhan Xu, Chun-Hao Huang, Krishna Kumar Singh, Yang Zhou, Leonidas J Guibas, and Jimei Yang. Actanywhere: Subject-aware video background generation. *Advances in Neural Information Processing Systems*, 37:29754–29776, 2024.
- [54] Yuchao Gu, Yipin Zhou, Bichen Wu, Licheng Yu, Jia-Wei Liu, Rui Zhao, Jay Zhangjie Wu, David Junhao Zhang, Mike Zheng Shou, and Kevin Tang. Videoswap: Customized video subject swapping with interactive semantic point correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7621–7630, 2024.
- [55] Sanoojan Baliah, Qinliang Lin, Shengcui Liao, Xiaodan Liang, and Muhammad Haris Khan. Realistic and efficient face swapping: A unified approach with diffusion models. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1062–1071. IEEE, 2025. 3
- [56] Xiangyang Luo, Ye Zhu, Yunfei Liu, Lijian Lin, Cong Wan, Zijian Cai, Yu Li, and Shao-Lun Huang. Canonswap: High-fidelity and consistent video face swapping via canonical space modulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10064–10074, 2025.
- [57] Hao Shao, Shulun Wang, Yang Zhou, Guanglu Song, Dailan He, Zhuofan Zong, Shuo Qin, Yu Liu, and Hongsheng Li. Vividface: A robust and high-fidelity video face swapping framework. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 3
- [58] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. Internvid: A large-scale video-text dataset for multimodal understanding and generation. In *The Twelfth International Conference on Learning Representations*, 2024. 3, 6
- [59] Nan Huang, Wenzhao Zheng, Chenfeng Xu, Kurt Keutzer, Shanghang Zhang, Angjoo Kanazawa, and Qianqian Wang. Segment any motion in videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3406–3416, 2025. 3, 5, 6
- [60] Bojia Zi, Weixuan Peng, Xianbiao Qi, Jianan Wang, Shihao Zhao, Rong Xiao, and Kam-Fai Wong. Minimax-remover: Taming bad noise helps video object removal. *arXiv preprint arXiv:2505.24873*, 2025. 4, 5
- [61] Zhe Chen, Jiannan Wu, Wenhui Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024. 5
- [62] Xun Long Ng, Kian Eng Ong, Qichen Zheng, Yun Ni, Si Yong Yeo, and Jun Liu. Animal kingdom: A large and diverse dataset for animal behavior understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19023–19034, 2022. 5, 13
- [63] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018. 5, 13
- [64] Lingyi Hong, Wenchao Chen, Zhongying Liu, Wei Zhang, Pinxue Guo, Zhaoyu Chen, and Wenqiang Zhang. Lvos: A benchmark for long-term video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13480–13492, 2023. 5, 13
- [65] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 724–732, 2016. 5, 13
- [66] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 5
- [67] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022. 6

- [68] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Kopula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier J Henaff, Matthew Botvinick, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. Perceiver IO: A general architecture for structured inputs & outputs. In *International Conference on Learning Representations*, 2022. [6](#)
- [69] Prolific. Prolific — easily collect high-quality data from real people, 2025. [7](#)
- [70] Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. [7](#)

## Appendix

This appendix includes additional materials that cannot fit into the main paper due to space constraints, and is structured as follows.

- Sec. A: Detailed specifications of the StM-50K dataset;
- Sec. B: Additional qualitative results and comparisons;
- Sec. C: Further discussion and details of StM inference efficiency and limitations.

### A. StM-50K Dataset Details

**Data Sources and Composition.** The StM-50K multi-layer video dataset comprises approximately 50,000 video clips and was constructed using our automated Decomposer pipeline (Section 3.3). For large, unannotated sources like Panda-70M [33] and Animal Kingdom [62], the full Decomposer pipeline—including motion segmentation and video inpainting—was utilized to generate the foreground, mask, background, and caption layers. For existing video object segmentation datasets (*e.g.*, YouTube-VOS [63] and LVOS [64]), we adapted the pipeline to leverage their provided ground-truth foreground masks, thus ensuring high-fidelity layer separation. The DAVIS dataset [65] was reserved exclusively for model validation during training.

To ensure temporal consistency and support multiple temporal resolution across the dataset, we apply a randomized preprocessing strategy. Videos are either directly split into multiple chunks of 49 frames, or first downsampled in frame rate (by a factor of 2, 3, or 4) and subsequently split into 49-frame chunks. This approach guarantees uniform input dimensions while preserving motion diversity.

**Test Benchmark.** For the final evaluation, we curated a dedicated test benchmark of 93 unique, unseen triplets (foreground video, background video, text prompt) from a held-out set. Crucially, the foreground and background components for each test sample were intentionally sourced from different original videos. This rigorous design robustly measures the model’s ability to generalize and synthesize plausible, novel interactions between disparate visual elements. The source videos for this benchmark were primarily selected from the DAVIS dataset [65], with additional samples sourced from the Animal Kingdom dataset [62].

### B. Additional Qualitative Results and Instructions

**User Study Details.** Our pairwise preference study involved 50 subjects recruited via the Prolific platform. The evaluators performed a total of 25 pairwise comparisons, randomly sampled from our test dataset, and were asked to select the superior output video (StM vs. a baseline) based on five distinct criteria.

The user study procedure, illustrated by the interface layout in Figure 9, began with an instructional phase (Figures 9a and 9b). This phase explained the required task inputs and provided an example comparison. Evaluators then proceeded to the evaluation phase, which consisted of a series of 25 questions. As shown in Figures 9c and 9d, each question page displayed the reference context, metric definitions, and the side-by-side video comparison for rating. To mitigate bias, the left-right presentation order of the compared methods was randomly shuffled for each question. This question format is repeated for all 25 samples for every baseline comparison.

The criteria used were:

- **Identity Preservation (FG & BG):** Measures visual identity fidelity of the foreground subject and background scene.
- **Motion Alignment (FG & BG):** Assesses the fidelity of subject and camera/scene motion.
- **FG-BG Harmony:** Assesses plausible interaction and affordance-awareness.
- **Overall Quality:** A holistic judgment of realism, artifacts, and coherence.

**VLLM-as-a-Judge Instructions.** We utilized **Gemini 2.5 Pro** (Public version) as a Vision-Language Large Model (VLLM) to act as an automated judge on the full test dataset. The VLLM was provided with the same pairwise comparison task as the human subjects. To ensure fair evaluation, we employed the specific system instructions listed below for each metric. The model was instructed to output a single character: A, B, or N (No Preference).

#### a. Foreground Identity Consistency:

*”You are a video analysis tool. You will be given a Reference Foreground video and two generated videos (Video A, Video B). IMPORTANT: Judge \*only\* this specific metric. Do not let overall visual quality influence your choice. Metric: Foreground Identity Consistency. Question: Which generated video (Video A or Video B) better preserves the appearance of the subject (person, animal, object) from the Reference Foreground video?”*

#### b. Foreground Motion Consistency:

*”You are a video analysis tool. You will be given a Reference Foreground video and two generated videos (Video A, Video B). IMPORTANT: Judge \*only\* this specific metric. Do not let overall visual quality influence your choice. Metric: Foreground Motion Consistency. Question: Which generated video (Video A or Video B) better preserves the subject’s motion from the Reference Foreground video and looks more physically believable?”*

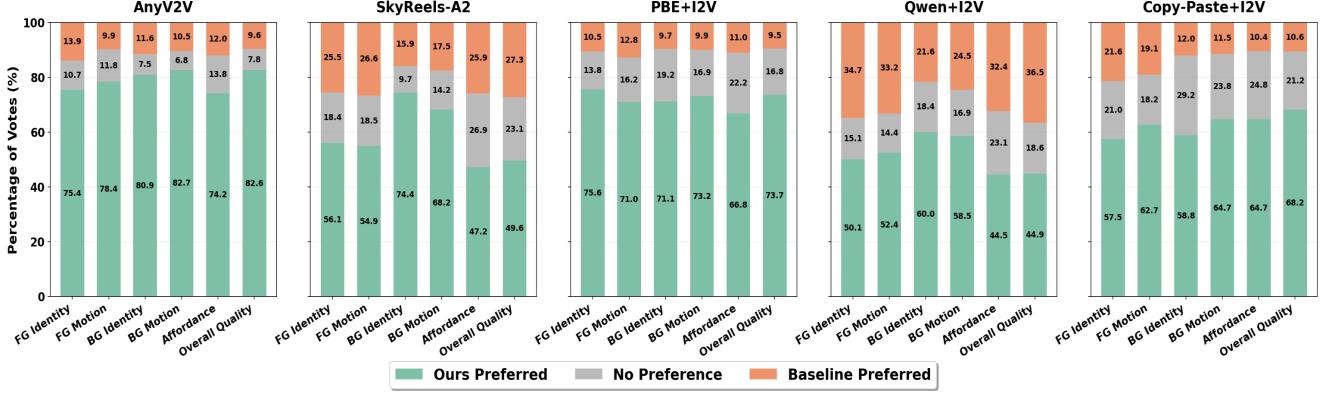


Figure 7. **User Study Results.** Pairwise preference win/tie/lose rates comparing our method (StM) against five baselines across six criteria. Preference for StM is shown in Green, ties in Grey, and preference for the baseline in Orange. The results demonstrate that StM consistently outperforms all baselines, achieving its highest preference rates in motion and identity preservation.

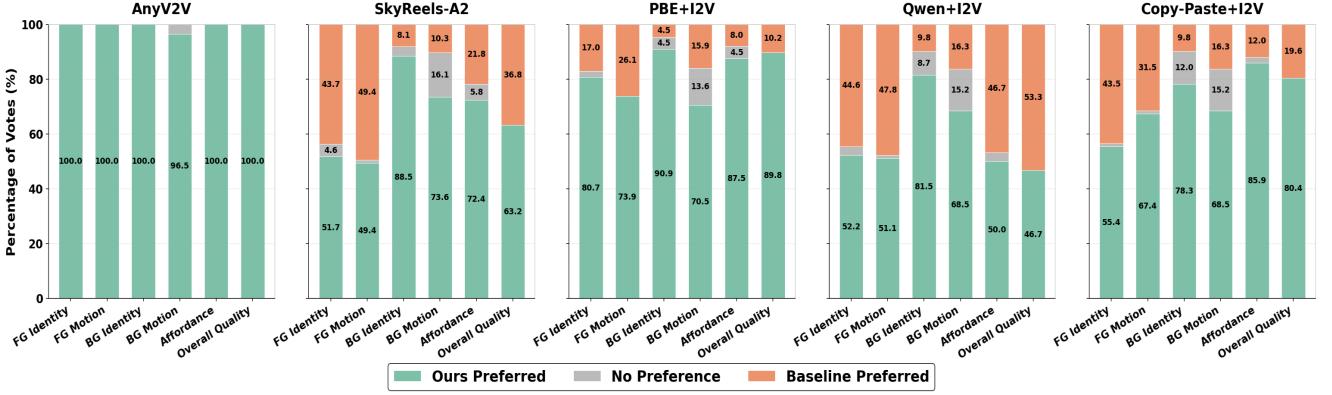


Figure 8. **VLLM-as-a-Judge Results.** Results from our automated evaluation using Gemini 2.5 Pro, which follows the exact pairwise comparison protocol employed in the human study. The VLLM judge demonstrates a strong alignment with human preferences, consistently favoring our StM in key motion and identity metrics.

### c. Background Identity Consistency:

*"You are a video analysis tool. You will be given a Reference Background video and two generated videos (Video A, Video B). IMPORTANT: Judge \*only\* this specific metric. Do not let overall visual quality influence your choice. Metric: Background Identity Consistency. Question: Which generated video (Video A or Video B) better preserves the appearance of the background scene from the Reference Background video?"*

### d. Background Motion Consistency:

*"You are a video analysis tool. You will be given a Reference Background video and two generated videos (Video A, Video B). IMPORTANT: Judge \*only\* this specific metric. Do not let overall visual quality influence your choice. Metric: Background Motion Consistency. Question: Which generated video (Video A or Video B) better preserves the background's motion (or camera movement) from the Reference Background video, and which looks smoother?"*

### e. Affordance-aware Generation:

*"You are a video analysis tool. You will be given a Reference Foreground, a Reference Background, and two generated videos (Video A, Video B). IMPORTANT: Judge \*only\* this specific metric. Do not let overall visual quality influence your choice. Metric: Affordance-aware Generation. Question: Which generated video (Video A or Video B) shows a more believable interaction between the subject and the background? (e.g., sitting \*on\* a chair, not \*through\* it; not walking through walls)."*

### f. Overall Quality:

*"You are a video analysis tool. You will be given a Reference Foreground, a Reference Background, and two generated videos (Video A, Video B). Metric: Overall Quality. Question: Which generated video (Video A or Video B) has the best overall visual quality, clarity, and fewest visual artifacts (like flickering, blurring, or blockiness)?"*

## Video Comparison Study (v1)

This is a research study for a **video composition** project. No personal information will be requested or collected in this survey. You will see 25 examples.

**How this study works:**

[Sign in to Google](#) to save your progress. [Learn more](#)

On each page, you will first see a **Prompt / Context** video (like the one below). This prompt shows the inputs:

1. The **Foreground Subject** (e.g., the person or animal).
2. The **Background Scene**.
3. The **Text Prompt** describing the desired action.

1. Example: Prompt / Context GIF (Shows Inputs)

Foreground Video      Background Video      Text Prompt

Below the prompt, you will see two generated videos, **Video A (Left)** and **Video B (Right)** (like the one below). These are two different 'compositions' of the foreground subject into the background scene.

(a) Instructions: Study Overview & Inputs

## Video Comparison Study (v1)

[Sign in to Google](#) to save your progress. [Learn more](#)

\* Indicates required question

**Example 1**

Please evaluate based on these metrics:

**IMPORTANT NOTE:** For the first 5 metrics, please try to judge only that specific metric. Do not let the overall visual quality influence your choice, except for the final metric ("6. Overall Quality").

**Metric Definitions (Grid Rows):**

- Foreground Identity Consistency:** Does the main subject (person, animal, object) in the generated videos (Video A or Video B) look similar to the subject in the Foreground Video?
- Foreground Motion Consistency:** Is the main subject's motion (e.g., walking) preserved from the input and physically believable?
- Background Identity Consistency:** Does the background / scene of the generated videos (Video A or Video B) look similar to the scene from the Background Video?
- Background Motion Consistency:** Is the background's motion (or camera movement) preserved from the input and does it look smooth and consistent?
- Affordance-aware Generation:** Does the subject interact with the background believably? (e.g., sitting on a chair, not through it; not walking through walls).
- Overall Quality:** Overall visual quality, clarity, and lack of visual artifacts (like flickering, blurring, or blockiness).

Prompt / Context

Foreground Video      Background Video      Text Prompt

(c) Question Interface: Context & Metrics

2. Example: Comparison GIF (Shows Results)

Video A      Video B

**Your Task:**  
Your task is to compare Video A and Video B based on the set of metrics provided on each page. For each metric, please choose the video that you believe performs better.

**IMPORTANT NOTE:** For the first 5 metrics, please try to judge \*only\* that specific metric. Do not let the overall visual quality (like blurriness or flickering) influence your choice, **except for the final metric ("6. Overall Quality")**.

*Note: The order of the generated videos (A vs. B) is randomly shuffled for each example to ensure fairness.*

[Next](#)      [Clear form](#)

(b) Instructions: Example Comparison

Comparison: Video A (Left) vs. Video B (Right)

Video A      Video B

Please evaluate the videos based on the metrics below: \*

	Video A (Left)	Video B (Right)	No Preference / Equal
1. <b>Foreground Identity Consistency</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
2. <b>Foreground Motion Consistency</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
3. <b>Background Identity Consistency</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
4. <b>Background Motion Consistency</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
5. <b>Affordance-aware Generation</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
6. <b>Overall Quality</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

[Back](#)      [Next](#)      [Clear form](#)

(d) Question Interface: Video Comparison & Rating

Figure 9. **User Study Interface.** (a)-(b) The initial instruction pages presented to the subject, defining the inputs and demonstrating the task. (c)-(d) An example of a single question page. This interface—showing the context, metric definitions, and side-by-side video comparison—is repeated 25 times, once for each randomly sampled test case in the study.

## C. Limitation and Efficiency Discussion

While StM achieves significant advances in affordance-aware video composition, we acknowledge inherent design trade-offs that present clear avenues for future work.

Our method is deliberately designed to prioritize the faithful preservation of input video motion and appearance as the dominant control signals, sometimes at the expense of strict semantic adherence to the text prompt. This design choice is evidenced by the quantitative results (Table 2 of the main paper), where text-guided baselines like SkyReels achieve higher Textual Alignment scores (M4) than StM (24.40 vs. 19.81). This suggests that for applications where semantic text control is paramount, StM may introduce minor visual compromises to maintain layer fidelity. Future work could address this by exploring a dynamic weighting scheme to better balance the influence of visual inputs and text guidance.

Furthermore, the overall performance relies heavily on the quality and reliability of the off-the-shelf Decomposer models (*e.g.*, motion segmentation and video inpainting). Errors introduced during foreground mask extraction or background inpainting can result in artifacts that the Composer subsequently inherits or struggles to fully correct. Improving the robustness of the decomposition phase remains a vital future direction.

The computational overhead for the StM framework is primarily concentrated during the *training phase* due to the transformation-aware training pipeline. Specifically, the model requires encoding three distinct video inputs ( $V_{org}$ ,  $V_{bg}$ ,  $\tilde{V}_{fg}$ ) through the frozen Space-Time VAE. The Composer model itself is based on a latent diffusion transformer and was fine-tuned for 20K iterations on 16 NVIDIA H100 GPUs with a total batch size of 64.

During Inference, the process is highly efficient. A key advantage of our design is that we strictly preserve the architecture of the base model (CogVideoX-I2V), adding only a lightweight projection layer at the input stage. Since this projection is negligible in size compared to the transformer backbone, our method incurs no additional computational complexity during the iterative denoising process. Consequently, the inference latency is effectively identical to the base I2V model, with the only minor overhead being the one-time encoding of the additional video layers.