Layer-Aware Video Composition via Split-then-Merge

Ozgur Kara[1†] Yujia Chen[2] Ming-Hsuan Yang[2]
James M. Rehg[1] Wen-Sheng Chu[2‡] Du Tran[2‡]

[1]University of Illinois Urbana-Champaign, USA [2]Google, USA

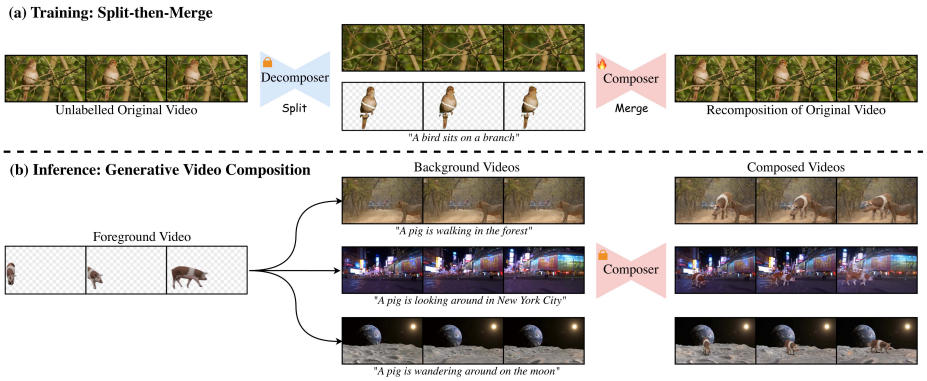

Fig. 1: Video Composition via Split-then-Merge. (a) Training: The Decomposer *splits* an unlabeled video into foreground and background layers and generates a caption, while the Composer learns to *merge* them for reconstruction. **(b) Inference:** The Composer integrates a foreground video into novel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

Abstract. We present **S**plit-**t**hen-**M**erge (StM), a controllable generative video composition framework that minimizes reliance on annotated datasets and handcrafted rules. Instead of requiring manual supervision, StM decomposes unlabeled videos into dynamic foreground and background layers. By self-composing these elements, the model learns to synthesize complex interactions between moving subjects and diverse scenes, capturing the intricate dynamics necessary for high-fidelity video generation. Specifically, StM introduces a transformation-aware training pipeline utilizing multi-layer fusion and augmentation to address affordance challenges in video composition. An identity-preservation loss further maintains foreground fidelity during blending. Extensive experiments show that StM outperforms state-of-the-art methods on both quantitative benchmarks and qualitative evaluations, including human studies and Vision-Language Model assessments. Finally, we release StM-50K, the first multi-layer video dataset, to facilitate future research in generative video composition. Data, code, and more details are available at our project page.

† Work done during an internship at Google.
‡ Joint last authors.

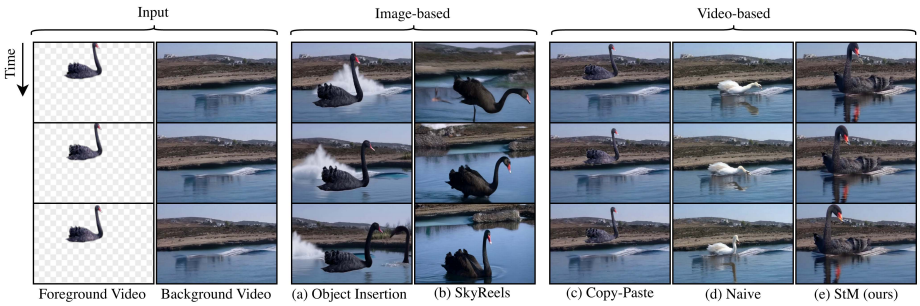

Fig. 2: Video Composition. Given input foreground and background videos, image-based methods (a)–(b) use only the first frame, while (c)–(e) take full video inputs. (a) Object insertion [59] followed by Image-to-Video (I2V) and (b) end-to-end I2V composition SkyReels [13] fails to retain motion due to 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).

1 Introduction

Recent advances in diffusion models [21, 44, 50, 53, 66] have greatly improved video realism. However, controllability remains largely limited to text [2, 16, 20, 23, 30, 65] and image conditioning [14, 18, 29, 40, 42, 49, 52, 68], which lack precision for precise composition. While structured signals like pose [5, 24, 31, 38] or motion [4, 15, 39] offer stronger control, most methods focus on animating static images rather than flexible synthesis. Professional content creation relies on layer-based compositing [3,58], motivating *generative video composition*: synthesizing a coherent video by integrating a dynamic foreground with a separate background using generative models, as depicted in Fig. 1(a). Unlike classical composition [6], which emphasizes matting and blending, the generative setting allows adaptive adjustments to foreground to harmonize with new lighting and camera dynamics. As shown in Fig. 1(b), the foreground animal's motion, appearance, and shadows are adapted to match diverse backgrounds.

One natural question is: "*How challenging is this task, and can it be addressed by existing methods?*" A straightforward baseline combines image-based object insertion with an image-to-video (I2V) model. Specifically, the first frames of the foreground and background videos are composed using a state-of-the-art insertion method [59], and the resulting image is provided to an I2V model [65] for video generation. As shown in Fig. 2(a), this pipeline discards the temporal information of the foreground video, requiring the I2V model to infer motion solely from text. More recent approaches, such as SkyReels [13], which directly compose videos from foreground and background images, exhibit a similar limitation due to the absence of explicit motion cues (Fig. 2(b)).

An ensuing question is "*Can this task be addressed by extending methods to video inputs?*" The answer is negative. A naive copy-and-paste strategy lacks affordance awareness and often yields implausible placements. As shown in Fig. 2(c),

the swan is incorrectly positioned on the ground rather than in water. Fine-tuning a state-of-the-art video generator [65] with sufficient data is another attempt. However, naively trained models remain prone to affordance violations due to shortcut learning and fail to preserve motion and appearance consistency. Fig. 2(d) shows the swan's appearance shifting to white, violating foreground identity preservation. These results indicate that generative video composition is non-trivial and requires dedicated designs beyond off-the-shelf adaptations.

In this paper, we propose **S**plit-**t**hen-**M**erge (StM), a data-driven framework for video composition requiring *no* manual annotation (Fig. 1). Our key idea is to decompose unlabeled videos into foreground-background layers and train the model to reconstruct them, enabling scalable self-composition on large datasets (*e.g.*, Panda-70M [7]). Unlike application-specific editing methods (*e.g.*, personalization [8], object swapping [17], motion control [4,15,39,54]), StM is fully data-driven and generalizes across objects given a reliable Decomposer. We introduce two key components: (1) *a transformation-aware training pipeline* that prevents shortcut learning and promotes affordance awareness, and (2) *an identity-preservation loss* that maintains foreground identity while ensuring seamless background integration. In addition, we propose quantitative metrics for appearance and motion consistency, and leverage Vision-Language Models (VLMs) as automated judges to complement human evaluation. Our key contributions are:

1. We propose StM, a scalable framework for generative video composition that eliminates manual annotations. StM includes two novel components: the transformation-aware training pipeline and the identity-preservation loss.
2. We release StM-50K, the first multi-layer video dataset created through an automatic pipeline, which enables training future layer-based video generation methods.
3. We present a comprehensive evaluation including strong baselines, new quantitative metrics, user studies, and VLM-based judges, demonstrate that StM substantially outperforms existing methods.

2 Related Work

Controllable Video Generation. The success of diffusion models in image synthesis [21,50,53] has driven rapid progress in video generation, shifting emphasis from visual quality to controllability. Text-guided models [2,16,20,23,30,32,65] provide high-level semantic guidance but lack precise control over identity and complex motion. Image-to-video (I2V) models [14,18,29,40,42,49,52,68] improve identity preservation via reference images, yet their static conditioning fails to capture temporal dynamics. Structural controls, such as pose [5,24,31,38] or spatial signals [19,40,55,60,61,67], offer fine-grained guidance but remain limited to image-level animation. Consequently, existing methods are inherently image-centric, controlling isolated attributes rather than full video dynamics. In contrast, we tackle affordance-aware composition that integrates identity and motion from a foreground video with the dynamics of a background video.

Image and Video Composition. Existing composition methods mainly rely on pixel-level blending techniques, such as color transfer [46,48], harmoniza-

tion [11], and gradient-domain pasting [28], to enhance realism [34, 63]. Video extensions incorporate motion cues [6, 56] but lack generative capabilities. Generative harmonization of two dynamic video layers remains largely unexplored. Common baselines first insert objects at the image level (*e.g.*, PbE [64], AnyDoor [9], Qwen-Edit [59]), then animate the result with an I2V model; SkyReels [13] follows a similar paradigm. However, these methods discard the original foreground motion by conditioning on static images. Other approaches differ in scope: AnyV2V [33] propagates single-image edits, and VideoAnyDoor [54] inserts a static image into a video. The most related work, LayerFlow [27], generates all layers from text. In contrast, StM composes two existing video layers, explicitly preserving both their identity and motion.

3 Video Composition via Split-then-Merge

3.1 Generative Video Composition

Generative video composition aims to synthesize a coherent video by integrating a subject—together with its appearance and motion—from one source video into the dynamic scene of another. Formally, the inputs include: (i) a foreground video $V_{fg} \in [0, 255]^{T\times H\times W}$ (omitting the color channel for simplicity), containing a primary subject and a binary mask $M_{fg} \in \{0, 1\}^{T\times H\times W}$ associated with it; (ii) a background video V_{bg} providing scene context; and (iii) a text prompt $\mathcal{T}$ describing the target scene. The objective is to generate V_{pred} that seamlessly integrates the subject from V_{fg} into V_{bg} while preserving visual and motion consistency.

This task poses two challenges. First, *layer consistency* requires preserving foreground appearance and motion (*e.g.*, subject dynamics) as well as background scene and camera motion to avoid pasted-on artifacts. Second, *affordance awareness* requires semantic understanding to ensure physically plausible placement and interaction (*e.g.*, placing a car 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 based on *self-composition*. The key idea is to use off-the-shelf models to *split* an unlabeled video V_{org} into layers and train a model to *merge* them for reconstruction (Fig. 1(a)). Without human annotation, the StM Decomposer automatically separates any unlabeled video into a foreground subject and background scene, and generates a text caption. This process converts raw videos into training samples, enabling large-scale multi-layer dataset construction. A diffusion-based generative model (*i.e.*, the Composer) is then trained to *merge* the layers back into the original coherent video.

This approach differs from conventional generative video and computational photography methods, which rely on specialized algorithms or domain heuristics [1,16,17,27,33,43]. Such methods are often confined to narrow domains (*e.g.*, face swapping [1, 37, 51]), limiting scalability to diverse in-the-wild content. By casting the task as inverting a decomposition process, our framework avoids these constraints and learns realistic video composition directly from data.

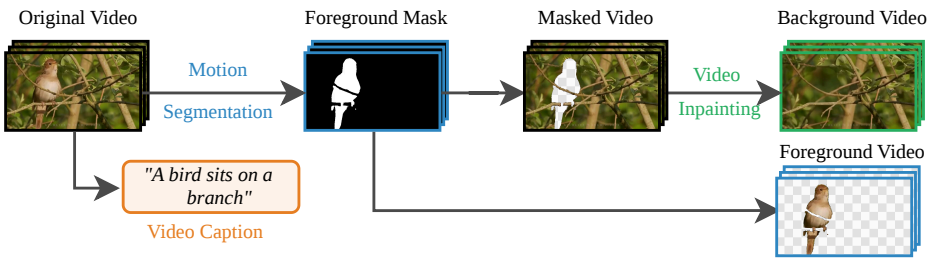

Fig. 3: StM Decomposer. The StM Decomposer integrates off-the-shelf models to split unlabeled videos. First, motion segmentation generates a foreground mask, which is used to extract the foreground layer. An inpainting model then fills the "holes" in the masked background video. Finally, a video captioning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

3.3 Decomposer: Splitting Videos into Layers

A key challenge in generative video composition is the lack of large-scale multi-layer video datasets. To address this, we design an automated pipeline that decomposes videos into four components: a caption, a foreground layer, its mask, and a background layer. As shown in Fig. 3, the StM Decomposer first employs a pre-trained video-language model [57] to generate captions. An automatic motion segmentation model (*i.e.*, Segment-Any-Motion [25]) then extracts the primary moving subject and its mask. The masked region is removed, and a state-of-the-art video inpainting model [69] reconstructs a coherent background.

3.4 Composer: Learning to Merge

The StM Composer builds upon a latent diffusion transformer [65] for text-to-video generation. It consists of a pre-trained space-time VAE encoder-decoder (E_v, D_v), a text encoder (E_T), and a Diffusion Transformer (DiT) backbone $\mathbf{f}(x_t; \theta)$, where x_t denotes the latent input, θ represents the model parameters, and t is the diffusion timestep. The standard objective is to recover the clean latent from a noise-added latent conditioned on text embeddings. We adapt this architecture for video composition via a transformation-aware training pipeline with three components: multi-layer conditional fusion, transformation-aware augmentation, and an identity-preservation loss.

Multi-layer Conditional Fusion. Fig. 4 illustrates the StM Composer architecture. The foreground video V_{fg} is first augmented to obtain $\tilde{V}_{\text{fg}}$. The original video, background video, and augmented foreground are encoded by E_v to produce latent representations $\mathbf{z}_{\text{org}}, \mathbf{z}_{\text{bg}}, \tilde{\mathbf{z}}_{\text{fg}}$, where $\mathbf{z}_i = E_v(V_i)$. During diffusion, noise ϵ_t is added to $\mathbf{z}_{\text{org}}$ following diffusion schedule $\bar{\alpha}_t$, and all visual latents are fused via channel-wise concatenation followed by an MLP:

$$\mathbf{z}_{\text{vision},t} = \text{MLP}(\text{Concat}_{\text{channel}}(\sqrt{\bar{\alpha}_t}\mathbf{z}_{\text{org}} + \sqrt{1-\bar{\alpha}_t}\epsilon, \tilde{\mathbf{z}}_{\text{fg}}, \mathbf{z}_{\text{bg}})). \tag{1}$$

This *channel-level conditioning* provides dense spatio-temporally aligned guidance, in contrast to token-based or cross-attention conditioning. By integrating

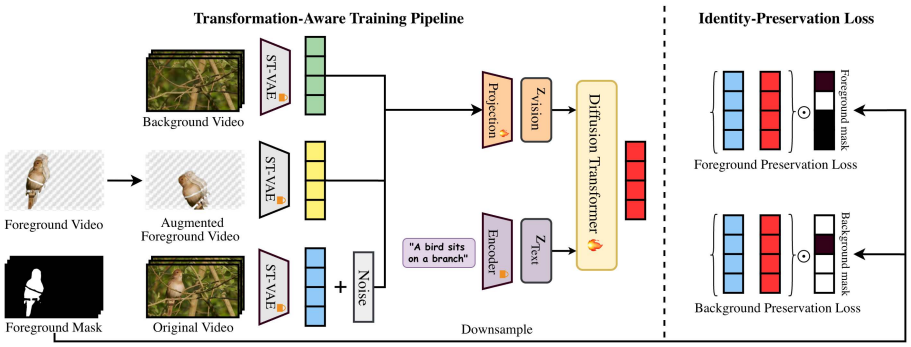

Fig. 4: StM Composer Training. The Composer is trained to reconstruct a ground-truth video latent from foreground, background, and text inputs. First, the foreground video is augmented, and all video inputs (augmented foreground, background, ground truth) are encoded into latents by a frozen Space-Time (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.

foreground, background, and noisy input at each latent location, it enables joint reasoning over all visual components and facilitates precise recomposition. The text prompt $\mathcal{T}$ is encoded as $\mathbf{z}_{\text{text}} = E_T(\mathcal{T})$. The final transformer input is $\mathbf{x}_t = \text{Concat}_{\text{token}}(\text{Tokens}(\mathbf{z}_{\text{vision},t}), \mathbf{z}_{\text{text}})$, which is fed into the DiT model to predict $\mathbf{z}_t = \mathbf{f}(\mathbf{x}_t; \theta)$, reconstructing the clean latent $\mathbf{z}_0 \triangleq \mathbf{z}_{\text{org}}$.

Transformation-Aware Augmentation. To prevent "copy-paste" shortcuts and promote genuine compositional reasoning, we introduce a targeted augmentation strategy. The key idea is to increase training difficulty by applying random transformations solely to the conditioning foreground V_{fg}. The model must therefore ***invert*** these transformations to correctly and plausibly place the subject within the background. Specifically, we apply spatial augmentations (random horizontal flipping, cropping, and resizing) and photometric changes (color jittering) to the entire video. By perturbing the subject's orientation, scale, position, and color, we prevent memorization of fixed layouts and encourage true ***affordance awareness*** and ***color harmony***. This compels the model to learn placement and harmonization principles rather than replicate patterns.

Identity-Preservation Loss. The standard latent diffusion objective is:

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

where we omit the constant weight w_t for simplicity. This loss treats all latent locations uniformly, as E_v preserves spatio-temporal locality and the ℓ_2 objective assigns equal weight across regions. However, video composition must balance foreground identity with visual harmony. Over-optimizing for harmony can dis-

tort identity (as shown in Fig. 2(d)), whereas strictly enforcing it often yields disjoint compositions.

We introduce an identity-preservation loss to balance foreground identity retention and overall harmony. Given the training-time foreground mask $\mathbf{M}$, we partition the latent space into foreground and background regions with separate weights, yielding:

$$\mathcal{L}_{\text{fg}} = \mathbb{E}_{\mathbf{z}_0,t}\left[\frac{\sum\left((\mathbf{z}_0 - \mathbf{f}(\mathbf{x}_t, t; \theta))^2 \odot \mathbf{M}\right)}{\sum \mathbf{M}}\right],$$
$$\mathcal{L}_{\text{bg}} = \mathbb{E}_{\mathbf{z}_0,t}\left[\frac{\sum\left((\mathbf{z}_0 - \mathbf{f}(\mathbf{x}_t, t; \theta))^2 \odot (1 - \mathbf{M})\right)}{\sum(1 - \mathbf{M})}\right]. \tag{3}$$

Both terms are normalized by their respective areas to reduce sensitivity to object size. The final loss is a weighted combination:

$$\mathcal{L}_{\text{final}} = \alpha\mathcal{L}_{\text{fg}} + (1 - \alpha)\mathcal{L}_{\text{bg}}. \tag{4}$$

This formulation explicitly controls the trade-off between identity preservation and compositional harmony.

Inference. Inference follows the training procedure with three differences: (1) the ground-truth latent is replaced by a sampled noise latent ϵ, (2) foreground augmentation is disabled, and (3) the space-time decoder (D_v) maps the predicted latent back to pixel space.

4 Experiments

4.1 Implementation Details

Decomposer. Our Decomposer (Section 3.3) processes unlabeled videos with off-the-shelf models: InternVL [10] for captioning, Segment-Any-Motion [25] for foreground masks, and MiniMax Remover [69] for background inpainting.

Constructing StM-50K dataset. Because there are no existing large-scale video datasets with comprehensive multi-layer annotations, creating this resource required rigorous curation and strategic sampling across diverse sources to balance semantic categories and capture a wide spectrum of complex motion dynamics. For training, we purposefully sourced videos from multiple datasets to encompass a diverse set of subjects: animals (Animal Kingdom [41]), humans (Panda-70M [7]), and general objects (Youtube-VOS [62], LVOS [22]). We then applied our automatic pipeline to decompose these sampled videos into layers, resulting in StM-50K, a training dataset of 50,000 processed clips.

For evaluation, we rely exclusively on the DAVIS dataset [45]. We curated 93 unseen triplets (foreground, background, text) where the foreground and background originate from different videos. These triplets span diverse semantic categories—including animals, human actions, and rigid objects—to rigorously test compositional generalization and ensure a comprehensive evaluation across varying motion complexities. Note this evaluation set size is comparable to standard video generation benchmarks [13, 27, 33] ranging from 50–100 videos.

Table 1: ***Quantitative Evaluation Metrics.*** Each metric computes the similarity or distance $f(.,.)$ between the representation of the ground truth ($\mathbf{r}_{\text{gt}}$) and the corresponding predicted ($\mathbf{r}_{\text{pred}}$) layers. Here $\mathbf{v}_{\text{gt}}$ and $\mathbf{v}_{\text{pred}}$ denote the ground truth and generated output videos, respectively, with superscripts FG and BG indicating the segmented foreground and background layers, and $\mathbf{t}$ denotes the guidance text prompt. ViCLIP [57], VideoSwin [35], and OptFlow [26] denote methods mapping video layers (or text) into representations.

Metric	$f(\cdot,\cdot)$	$\mathbf{r}_{\text{gt}}$	$\mathbf{r}_{\text{pred}}$
(M1) FG ID Preservation Consistency of the FG subject.	CosSim ↑	ViCLIP($\mathbf{v}_{\text{gt}}^{\text{FG}}$)	ViCLIP($\mathbf{v}_{\text{pred}}^{\text{FG}}$)
(M2) BG ID Preservation Consistency of the BG scene.	CosSim ↑	ViCLIP($\mathbf{v}_{\text{gt}}^{\text{BG}}$)	ViCLIP($\mathbf{v}_{\text{pred}}^{\text{BG}}$)
(M3) Semantic Action Alignment Fidelity of FG action to the ground-truth.	KL Div. ↓	VideoSwin($\mathbf{v}_{\text{gt}}^{\text{FG}}$)	VideoSwin($\mathbf{v}_{\text{pred}}^{\text{FG}}$)
(M4) BG Motion Alignment Consistency of BG motion dynamics.	MSE ↓	OptFlow($\mathbf{v}_{\text{gt}}^{\text{BG}}$)	OptFlow($\mathbf{v}_{\text{pred}}^{\text{BG}}$)
(M5) Textual Alignment Alignment between video and text.	CosSim ↑	ViCLIP($\mathbf{t}$)	ViCLIP($\mathbf{v}_{\text{pred}}$)

Composer. We adopt CogVideoX-I2V [65] as the base model, initialize with pre-trained weights, and fine-tune for 20K iterations. Training uses 16 NVIDIA H100 GPUs (batch size 64) with `bf16` precision. We employ AdamW [36] ($\beta_1 = 0.9$, $\beta_2 = 0.95$, $\epsilon = 1\text{e}{-8}$, weight decay 1e−4, max grad norm 1.0) and a cosine scheduler (base LR 5e−6, 1K warm-up). Videos are resized to 49×480×720, and the identity-preservation weight is $\alpha = 0.5$.

To mitigate shortcut learning, we apply the following temporally consistent augmentations sequentially to the foreground video: (1) random horizontal flipping ($p = 0.7$); (2) random resized cropping ($p = 0.7$, scale [0.5, 2.0], preserving 90% foreground); (3) color jittering ($p = 0.2$) exclusively on the masked foreground to perturb brightness, contrast, saturation, and hue within [0, 0.2].

4.2 Evaluation Setups

Baselines. We evaluate StM against 7 SoTA methods in three groups: *cascaded I2V*, *video-centric*, and *layer-aware* methods. To ensure a strictly fair comparison, cascaded I2V baselines utilize the exact same pre-trained CogVideoX-I2V backbone, inference steps, and classifier-free guidance scales as our StM Composer. These include **Copy-Paste + I2V** (centered paste with half-scale bounding box), **PBE + I2V** using Paint-by-Example [64], and **Qwen + I2V** using Qwen-Edit [59]. Video-centric baselines include **SkyReels** [13], which composes static images end-to-end, and **AnyV2V** [33], a tuning-free motion-aware editor. Layer-aware methods like **LayerFlow-FG** and **LayerFlow-BG** [27] condition on foreground or background layers, respectively. Notably, cascaded I2V baselines and SkyReels cannot preserve original foreground motion due to reliance on static inputs.

Metrics. We evaluate all methods using five criteria through automated metrics and qualitative studies (user study and VLM-as-a-Judge). For quantitative evaluation, we use [25] to decompose generated videos into foreground (FG) and

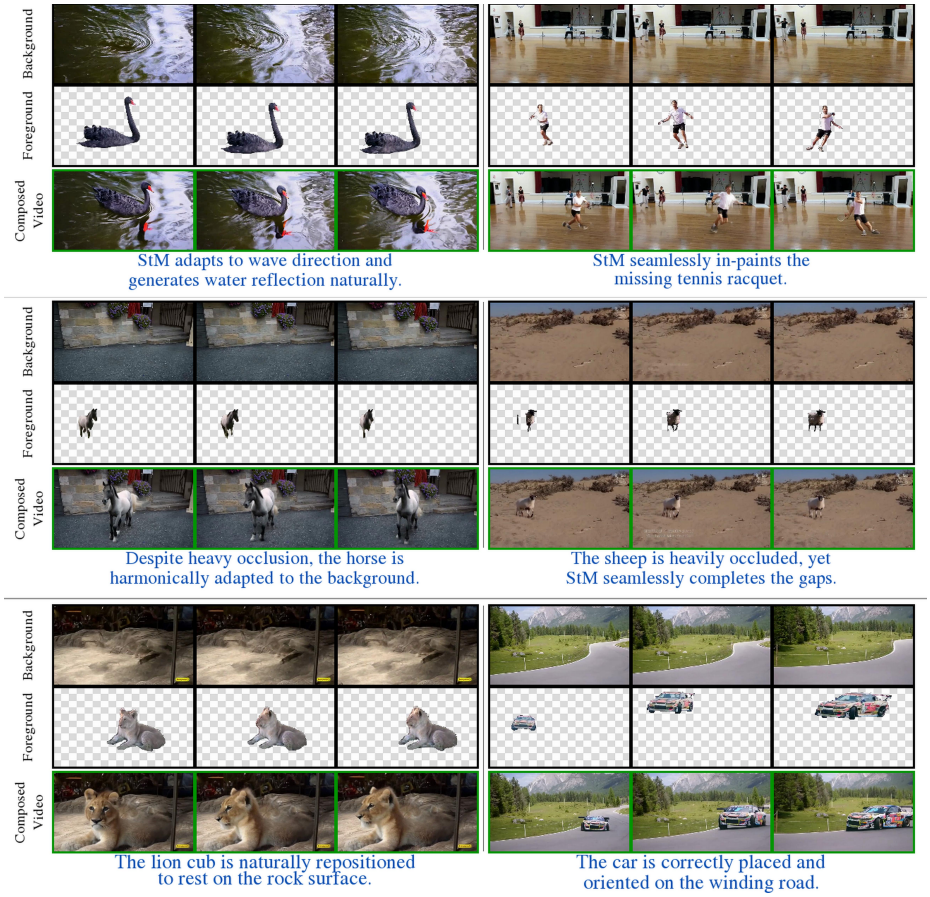

Fig. 5: Qualitative Results. StM demonstrates robust motion preservation and affordance awareness. Characteristic actions are faithfully retained (*e.g.*, cheetah, tennis player). Subjects are plausibly integrated rather than "pasted," as seen with the swan and duck adapted to water flow, and the lion and race car harmonized with appropriate scene lighting and placement. See Supplementary for full videos.

background (BG) layers, which are compared with input FG/BG layers using the metrics in Table 1. M1/M2 (FG/BG ID Preservation) measure appearance consistency; M3 (Semantic Action Alignment) evaluates FG semantic motion; M4 (BG Motion Alignment) measures BG pixel-level motion (camera and scene dynamics); M5 (Textual Alignment) measures fidelity of the composite video to the guidance text prompt. M1–M4 are purposely designed to evaluate video composition by directly measuring fidelity. M5 is a standard metric for image and video generation, which we provide for completeness. For qualitative evaluation, we conduct a user study with 50 participants on Prolific [47] using 25 randomly sampled test cases, and employ Gemini 2.5 Pro [12] as a VLM judge on the full test set (see Supplementary). The four criteria are: (i) ***Identity Preservation***

Table 2: Quantitative Evaluation and Ablation Study. Comparison with baseline methods (top) and ablations of each component (bottom) by removing them from our full model. We report **M1** (FG ID Preservation), **M2** (BG ID Preservation), **M3** (Semantic Action Alignment), **M4** (BG Motion Alignment), and **M5** (Textual Alignment). Metric directions are marked (↑, ↓). M1, M2, and M5 scores are multiplied by 100. **Bold** indicates the best performance, and underline is the second best. Percentage changes in the ablation study are relative to our full model.

Method		ID Preservation		Action / Motion Align.		Text Align.
		FG (M1) ↑	BG (M2) ↑	Sem. Action (M3) ↓	BG Motion (M4) ↓	(M5) ↑
Copy-Paste + I2V		83.08	85.02	1.61	184.59	19.21
PBE + I2V		73.52	80.19	2.47	98.08	19.41
Qwen + I2V		82.02	72.38	1.71	74.77	24.20
SkyReels		80.24	75.24	1.75	279.23	**24.40**
AnyV2V		77.73	56.36	1.70	154.97	24.13
LayerFlow-FG		83.02	60.21	1.47	143.47	18.74
LayerFlow-BG		50.27	91.73	2.35	24.72	19.57
StM (ours)		**84.82**	**92.88**	**1.22**	**16.36**	19.81
Aug.	**ID Loss**	**Ablation Study**				
✓	✓	84.82	**92.88**	1.22	**16.36**	**19.81**
✓	×	82.01 (-3.3%)	88.25 (-5.0%)	0.92 (-24.6%)	23.75 (+45.2%)	16.42 (-17.1%)
×	✓	83.15 (-2.0%)	89.20 (-4.0%)	0.70 (-42.6%)	16.61 (+1.5%)	16.56 (-16.4%)
×	×	**90.39** (+6.6%)	84.76 (-8.7%)	**0.77** (-36.9%)	18.59 (+13.6%)	18.70 (-5.6%)

(FG & BG); (ii) ***Motion Alignment (FG & BG)***; (iii) ***FG-BG Harmony*** (qualitative only); and (iv) ***Overall Quality*** (qualitative only).

4.3 Analysis

Quantitative Evaluation. Table 2 shows the quantitative advantages of StM. For ***Identity Preservation***, StM achieves the highest scores for both foreground (M1: 84.82) and background (M2: 92.88), indicating reduced appearance drift compared to I2V baselines that infer dynamics from a single frame (*e.g.*, PBE + I2V). The gap widens in ***Action & Motion Alignment*** (lower is better). StM attains the lowest FG action error (M3: 1.22), outperforming baselines (1.47–2.47) that hallucinate motion. For BG motion alignment, our method scores M4: 16.36, substantially better than competitors (24.72–279.23), which often fail to preserve camera and scene dynamics and generate static backgrounds.

For ***Textual Alignment*** (M5), our method scores 19.81, performing similarly to layer-aware baselines LayerFlow (18.74–19.57). This is expected, as StM relies primarily on foreground-background layers for control, with the text prompt provides complementary guidance to harmonize the composite rather than driving generation from scratch. Methods that achieve higher M5 scores (*e.g.*, SkyReels, Qwen + I2V) operate in a fully text-driven regime where the prompt is the sole generative signal, naturally yielding stronger text alignment at the cost of identity and motion fidelity. In addition, because we introduce *no new parameters beyond a projection layer*, our inference time remains identical to the CogVideoX-I2V baseline. These quantitative results align with qualitative fail-

Table 3: Pairwise Preference Study (Win Rates). We present a qualitative study where human evaluators and a VLM judge (Gemini 2.5 Pro) performed pairwise comparisons between our method and a baseline. The percentages reflect our method's "win rate" (*i.e.*, how often it was preferred over the baseline). Values in green indicate our method is preferred (>50%), while **red** indicates the baseline is preferred (<50%).

Ours vs. Baseline	User Study (in %)						VLM Judge (in %)					
	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		
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
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
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**
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
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
LayerFlow-FG	51.00	94.79	54.15	91.55	86.76	88.71	52.04	100.00	62.24	95.92	87.76	90.82
LayerFlow-BG	94.89	83.08	94.32	84.12	93.36	95.89	93.88	56.12	96.94	77.55	97.96	95.92

ures observed in baselines, such as a lack of affordance awareness in Copy-Paste + I2V and content inconsistency in Qwen + I2V.

User Study / VLM as a Judge. We conduct qualitative preference studies (Table 3) to complement automated evaluations (Table 2) and assess attributes beyond quantitative metrics, including FG-BG Harmony and Overall Quality. For Identity Preservation, both human and VLM judgments align with automated results, consistently favoring StM. For Action & Motion Alignment (FG & BG), StM achieves notably high win rates, highlighting a key limitation of I2V baselines that discard temporal dynamics by operating on static frames. In contrast, conditioning on full video layers enables StM to preserve motion. It is also strongly preferred for FG-BG Harmony (often >85%), demonstrating effective affordance-aware integration. High Overall Quality scores further reflect these advantages. Although Qwen + I2V is competitive in perceived quality, it performs poorly on motion and identity metrics, indicating that visual appeal alone does not ensure dynamic or identity fidelity. Appendix A2.2 (in the supplementary material) offers a focus evaluation with a greater emphasis on affordance awareness.

Qualitative Evaluation. Fig. 5 and 6 demonstrate StM's ability to produce affordance-aware, dynamically coherent compositions. As shown in Fig. 6, StM preserves rapid camera motion with a running goat (Left), correctly orients and positions a boat with wave dynamics (Center), and maintains a car's complex turn while aligning it with road geometry and lighting (Right). Fig. 5 further illustrates robust identity and motion preservation (cheetah, tennis player), alongside realistic integration of subjects (swan, duck, lion) with water flow, shadows, and illumination, avoiding a pasted appearance.

Ablation Study. Table 2 presents ablation results. The *w/o Both* variant learns a harmful copy-and-paste shortcut, yielding inflated FG scores (90.39 FG Identity, 0.77 FG Action) but poor compositional quality (18.59 BG Motion). Adding transformation-aware augmentation disrupts this shortcut: without *ID Loss*, FG Identity drops to 82.01, and FG Action slightly increases to 0.92,

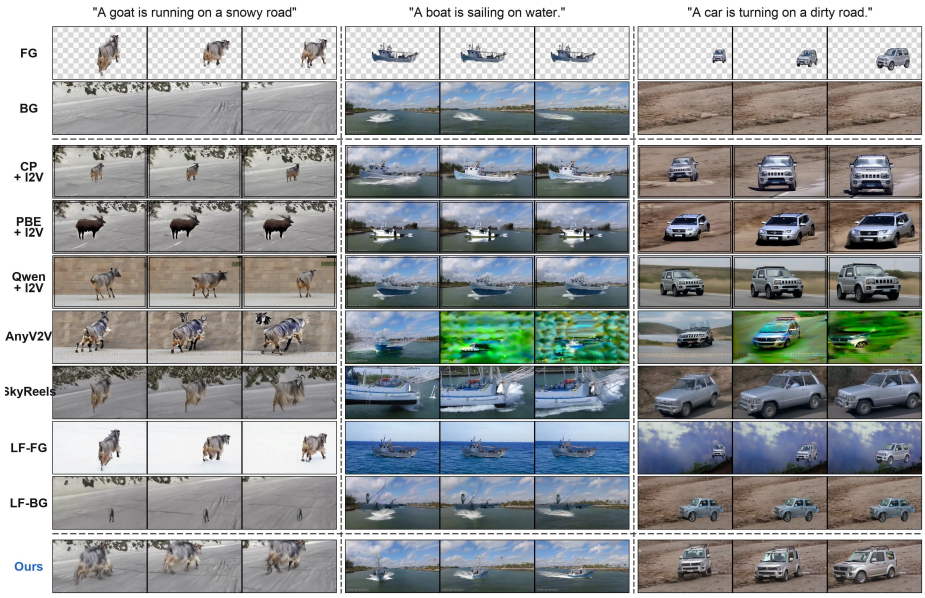

Fig. 6: Qualitative comparison. Our method (StM) uniquely preserves complex dynamics and achieves affordance-aware harmony where baselines fail. **(Left)** StM alone maintains both the rapid background camera motion and realistic foreground running motion. **(Center)**. StM demonstrates affordance by 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. See Supplementary for full videos.

reflecting genuine composition. Removing *Augmentation* alone fails to prevent shortcut learning (0.70 FG Action). The full **StM** balances both components, and achieves lower FG scores (84.82 FG Identity, 1.22 FG Action) but better BG Identity (92.88) and BG Motion (16.36). This confirms the need for both modules in robust, affordance-aware composition.

4.4 In-the-Wild Evaluation

We further evaluate StM through a series of *stress tests* along three dimensions: (i) generalization to real-world background videos with diverse scenes and lighting; (ii) logically inconsistent composition prompts; and (iii) iterative multi-object insertion. Foregrounds are drawn from the StM test set, while backgrounds are handheld videos captured in varied environments (*e.g.*, parks, backyards, living rooms) under changing illumination.

Generalization to Real-World Examples. Fig. 7(a) shows robust performance across scenes and lighting. StM seamlessly inserts an elephant into a night-time park, adapting illumination to ambient conditions, and reposes a pig to harmonize with balcony geometry and context.

Logically Impossible Composition. Fig. 7(b) presents results under semantically inconsistent prompts. StM re-orients a boat to align with road perspective

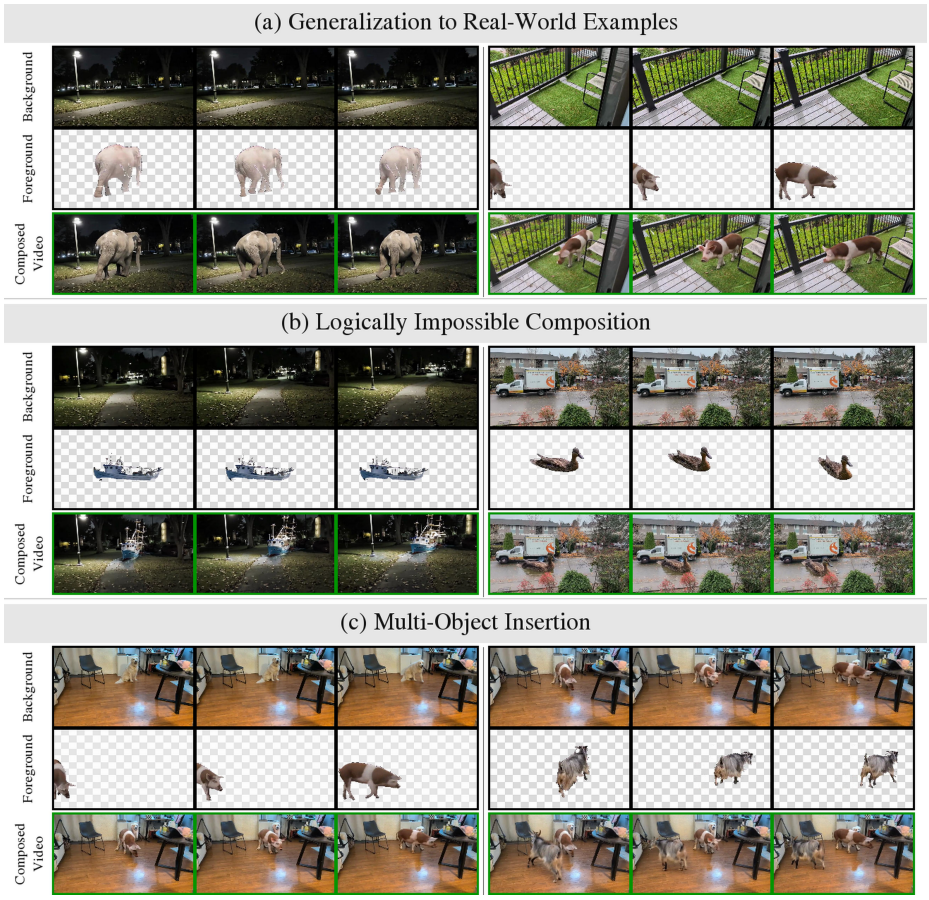

Fig. 7: StM In-the-Wild. StM is stress-tested to show its ability to: (a) generalize to smartphone videos with varying scenes and lighting, (b) creatively resolve logically impossible prompts such as "a boat floating on a road", and (c) seamlessly integrate multiple objects into a single background via recursive composition. Full-resolution videos and extended results are available in the Supplementary.

while synthesizing water on asphalt, and transforms part of a road into water to accommodate a swimming duck. These examples highlight strong geometric reasoning and affordance awareness.

Multi-Object Insertion. Fig. 7(c) demonstrates iterative composition, inserting a pig and a goat sequentially into a living room. StM grounds both subjects on the floor and synthesizes consistent shadows, indicating scalability to complex multi-object scenarios.

Challenging Test Cases. Fig. 8 illustrates challenging test cases for StM caused by severe occlusion and scale ambiguity. In the first example (Fig. 8(a)), StM cannot reconstruct the entire panda due to severe occlusion in the foreground input. However, it still manages to partially recover the subject and harmonize its col-

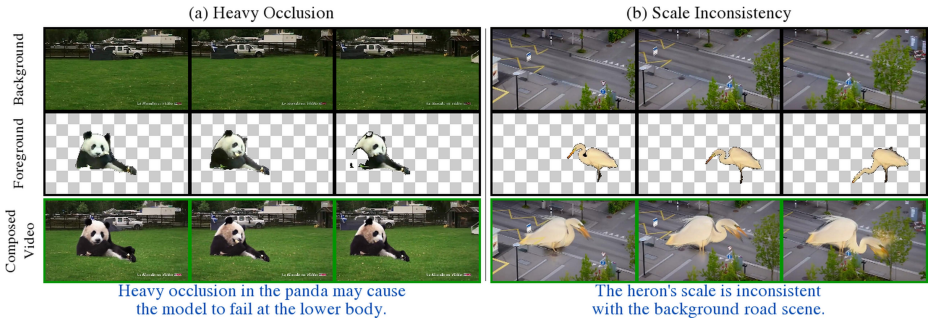

Fig. 8: Challenging Test Cases. (a) Severe occlusion in the foreground prevents full reconstruction of the panda. (b) Scale ambiguity from StM placing the heron with disproportionate size on the street. See Supplementary for videos.

ors with the background scene. In the second example (Fig. 8(b)), although StM plausibly grounds the heron on the street, the subject's scale is disproportionately large relative to the rest of the scene.

4.5 Robustness to Decomposer Quality

Because StM relies on an off-the-shelf Decomposer to produce its foreground (FG) layers, we evaluate its robustness to imperfect FG masks. We corrupt the FG masks during inference via two perturbations of increasing severity: *random pixel masking* (randomly removing interior pixels of the FG mask) and *boundary erosion* (removing pixels along the FG boundary). We then evaluate StM composition under these corrupted FGs (Table 4).

StM degrades gracefully, consistently outperforming LayerFlow-FG/BG across all metrics (M1–M4) and severities. It also surpasses AnyV2V entirely on M2 and M4, and maintains its lead on M1 (up to 25% masking and 50% erosion) and M3 (up to 10% masking and 25% erosion). This robustness is due to: (i) supervised losses operating in a latent space, making StM less sensitive to pixel variations; (ii) a highly diverse training dataset; and (iii) the use of imperfect FG inputs during training. Figure 9 shows StM composed results under different levels and types of corrupted FG.

Besides affecting inference, the Decomposer may also bias the evaluation metrics (M1–M4). We provide a full comparison of StM with other baselines using three different Decomposer options in Appendix A1.4 (in the supplementary material). The main findings still hold and are independent of the choice of Decomposer.

4.6 Limitation

StM still occasionally generates videos with scale inconsistencies between foreground objects and the background scene, *e.g.*, Fig. 8(b), the duck example in Fig. 7(b), and Fig. 2(e). It is inherently challenging for a purely data-driven approach to automatically deduce the relative sizes of disparate semantic categories—such as understanding that a duck is similar in size to a swan, but significantly

Table 4: Decomposer corruption (evaluated with SAM-3). StM composition under random pixel masking and boundary erosion of the foreground masks, compared against the clean setting and the video baselines.

	M1↑	M2↑	M3↓	M4↓	M5↑
Clean	82.22	90.60	1.06	16.88	19.81
LayerFlow-FG	53.99	26.35	3.97	176.24	18.74
LayerFlow-BG	51.93	80.15	4.12	79.13	19.57
AnyV2V	75.90	56.58	1.37	175.36	24.13
Rand. Pixel Masking 10%	79.59	89.83	1.29	17.38	19.26
Rand. Pixel Masking 25%	77.08	88.66	1.41	17.33	19.29
Rand. Pixel Masking 50%	73.27	87.07	1.72	26.94	18.47
Rand. Pixel Masking 75%	73.87	88.22	1.81	27.00	18.47
Mask Erosion 10%	79.09	89.08	1.30	17.20	19.42
Mask Erosion 25%	78.28	88.49	1.32	17.24	19.37
Mask Erosion 50%	76.86	87.40	1.44	19.35	19.11
Mask Erosion 75%	74.06	87.51	1.63	21.20	19.03

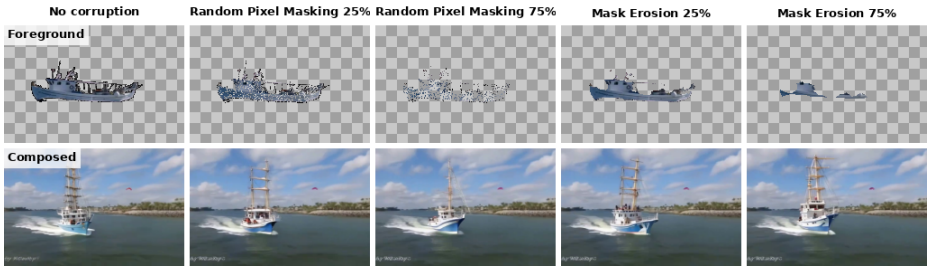

Fig. 9: StM composition under corrupted foregrounds. Randomly-masked and eroded foreground regions are shown for several severity levels; StM recovers a coherent composition in each case.

smaller than a boat. Future work will explore incorporating relative scale priors for semantic objects and scenes directly into the training pipeline.

5 Conclusion

We present StM, a practical pipeline for generative video composition that addresses data scarcity via a scalable "Split-then-Merge" strategy. By decomposing unlabeled videos into dynamic layers and self-composing them, StM learns subject-scene composition without manual annotation. Built on the StM-50K multi-layer dataset, our approach incorporates two key components: an *identity-preservation loss* for subject fidelity and *transformation-aware training* for motion and affordance modeling. Extensive experiments demonstrate that StM outperforms prior methods in motion preservation and compositional harmony.

References

1. Baliah, S., Lin, Q., Liao, S., Liang, X., Khan, M.H.: Realistic and efficient face swapping: A unified approach with diffusion models. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 1062–1071. IEEE (2025) 4
2. Bar-Tal, O., Chefer, H., Tov, O., Herrmann, C., Paiss, R., Zada, S., Ephrat, A., Hur, J., Liu, G., Raj, A., et al.: Lumiere: A space-time diffusion model for video generation. In: SIGGRAPH Asia 2024 Conference Papers. pp. 1–11 (2024) 2, 3
3. Brinkmann, R.: The art and science of digital compositing: Techniques for visual effects, animation and motion graphics. Morgan Kaufmann (2008) 2
4. Burgert, R., Xu, Y., Xian, W., Pilarski, O., Clausen, P., He, M., Ma, L., Deng, Y., Li, L., Mousavi, M., 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. pp. 13–23 (2025) 2, 3
5. Chang, D., Shi, Y., Gao, Q., Xu, H., Fu, J., Song, G., Yan, Q., Zhu, Y., Yang, X., Soleymani, M.: Magicpose: Realistic human poses and facial expressions retargeting with identity-aware diffusion. In: Forty-first International Conference on Machine Learning (2024), https://openreview.net/forum?id=jVXJdGQ4eD 2, 3
6. Chen, T., Zhu, J.Y., Shamir, A., Hu, S.M.: Motion-aware gradient domain video composition. Image Processing, IEEE Transactions on **22**(7), 2532–2544 (2013). https://doi.org/10.1109/TIP.2013.2251642 2, 4
7. Chen, T.S., Siarohin, A., Menapace, W., Deyneka, E., Chao, H.w., Jeon, B.E., Fang, Y., Lee, H.Y., Ren, J., Yang, M.H., 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. pp. 13320–13331 (2024) 3, 7
8. Chen, T.S., Siarohin, A., Menapace, W., Fang, Y., Lee, K.S., Skorokhodov, I., Aberman, K., Zhu, J.Y., Yang, M.H., Tulyakov, S.: Multi-subject open-set personalization in video generation (2025) 3
9. Chen, X., Huang, L., Liu, Y., Shen, Y., Zhao, D., Zhao, H.: Anydoor: Zero-shot object-level image customization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6593–6602 (2024) 4
10. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., 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. pp. 24185–24198 (2024) 7
11. Cohen-Or, D., Sorkine, O., Gal, R., Leyvand, T., Xu, Y.Q.: Color harmonization. ACM Transactions on Graphics **25**(3), 624–630 (2006) 4
12. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., 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) 9
13. Fei, Z., Li, D., Qiu, D., Wang, J., Dou, Y., Wang, R., Xu, J., Fan, M., Chen, G., Li, Y., et al.: Skyreels-a2: Compose anything in video diffusion transformers. arXiv preprint arXiv:2504.02436 (2025) 2, 4, 7, 8
14. Feng, W., Liu, J., Tu, P., Qi, T., Sun, M., Ma, T., Zhao, S., Zhou, S., HE, Q.: I2VControl-camera: Precise video camera control with adjustable motion strength. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=AcAD4VEgCX 2, 3

15. Geng, D., Herrmann, C., Hur, J., Cole, F., Zhang, S., Pfaff, T., Lopez-Guevara, T., Aytar, Y., Rubinstein, M., Sun, C., et al.: Motion prompting: Controlling video generation with motion trajectories. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 1–12 (2025) 2, 3
16. Geyer, M., Bar-Tal, O., Bagon, S., Dekel, T.: Tokenflow: Consistent diffusion features for consistent video editing. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=lKK50q2MtV 2, 3, 4
17. Gu, Y., Zhou, Y., Wu, B., Yu, L., Liu, J.W., Zhao, R., Wu, J.Z., Zhang, D.J., Shou, M.Z., Tang, K.: Videoswap: Customized video subject swapping with interactive semantic point correspondence. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7621–7630 (2024) 3, 4
18. Guo, X., Zheng, M., Hou, L., Gao, Y., Deng, Y., Wan, P., Zhang, D., Liu, Y., Hu, W., Zha, Z., et al.: I2v-adapter: A general image-to-video adapter for diffusion models. In: ACM SIGGRAPH 2024 Conference Papers. pp. 1–12 (2024) 2, 3
19. Guo, Y., Yang, C., Rao, A., Agrawala, M., Lin, D., Dai, B.: Sparsectrl: Adding sparse controls to text-to-video diffusion models. In: European Conference on Computer Vision. pp. 330–348. Springer (2024) 3
20. Gupta, A., Yu, L., Sohn, K., Gu, X., Hahn, M., Li, F.F., Essa, I., Jiang, L., Lezama, J.: Photorealistic video generation with diffusion models. In: European Conference on Computer Vision. pp. 393–411. Springer (2024) 2, 3
21. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020) 2, 3
22. Hong, L., Chen, W., Liu, Z., Zhang, W., Guo, P., Chen, Z., Zhang, W.: Lvos: A benchmark for long-term video object segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13480–13492 (2023) 7
23. Hong, W., Ding, M., Zheng, W., Liu, X., Tang, J.: Cogvideo: Large-scale pretraining for text-to-video generation via transformers. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=rB6TpjAuSRy 2, 3
24. Hu, L.: 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. pp. 8153–8163 (2024) 2, 3
25. Huang, N., Zheng, W., Xu, C., Keutzer, K., Zhang, S., Kanazawa, A., Wang, Q.: Segment any motion in videos. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 3406–3416 (2025) 5, 7, 8
26. Jaegle, A., Borgeaud, S., Alayrac, J.B., Doersch, C., Ionescu, C., Ding, D., Koppula, S., Zoran, D., Brock, A., Shelhamer, E., Henaff, O.J., Botvinick, M., Zisserman, A., Vinyals, O., Carreira, J.: Perceiver IO: A general architecture for structured inputs & outputs. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=fILj7WpI-g 8
27. Ji, S., Luo, H., Chen, X., Tu, Y., Wang, Y., Zhao, H.: 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. pp. 1–10 (2025) 4, 7, 8
28. Jia, J., Sun, J., Tang, C.K., Shum, H.Y.: Drag-and-drop pasting. ACM Transactions on Graphics **25**(3), 631–637 (2006) 4
29. Jiang, Y., Wu, T., Yang, S., Si, C., Lin, D., Qiao, Y., Loy, C.C., Liu, Z.: Videobooth: Diffusion-based video generation with image prompts. In: Proceedings of the

IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6689–6700 (2024) 2, 3
30. Kara, O., Kurtkaya, B., Yesiltepe, H., Rehg, J.M., Yanardag, P.: 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. pp. 6507–6516 (2024) 2, 3
31. Karras, J., Holynski, A., Wang, T.C., Kemelmacher-Shlizerman, I.: Dreampose: Fashion video synthesis with stable diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22680–22690 (2023) 2, 3
32. Khachatryan, L., Movsisyan, A., Tadevosyan, V., Henschel, R., Wang, Z., Navasardyan, S., Shi, H.: Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15954–15964 (2023) 3
33. Ku, M., Wei, C., Ren, W., Yang, H., Chen, W.: Anyv2v: A tuning-free framework for any video-to-video editing tasks. Transactions on Machine Learning Research (2024), https://openreview.net/forum?id=RFrJCkw2oa, reproducibility Certification 4, 7, 8
34. Lalonde, J.F., Efros, A.A.: Using color compatibility for assessing image realism. In: ICCV (2007) 4
35. Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H.: Video swin transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3202–3211 (2022) 8
36. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019), https://openreview.net/forum?id=Bkg6RiCqY7 8
37. Luo, X., Zhu, Y., Liu, Y., Lin, L., Wan, C., Cai, Z., Li, Y., Huang, S.L.: Canonswap: High-fidelity and consistent video face swapping via canonical space modulation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10064–10074 (2025) 4
38. Ma, Y., He, Y., Cun, X., Wang, X., Chen, S., Li, X., Chen, Q.: Follow your pose: Pose-guided text-to-video generation using pose-free videos. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 4117–4125 (2024) 2, 3
39. Mou, C., Cao, M., Wang, X., Zhang, Z., Shan, Y., Zhang, J.: Revideo: Remake a video with motion and content control. Advances in Neural Information Processing Systems **37**, 18481–18505 (2024) 2, 3
40. Namekata, K., Bahmani, S., Wu, Z., Kant, Y., Gilitschenski, I., Lindell, D.B.: SG-i2v: Self-guided trajectory control in image-to-video generation. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=uQjySppU9x 2, 3
41. Ng, X.L., Ong, K.E., Zheng, Q., Ni, Y., Yeo, S.Y., Liu, J.: Animal kingdom: A large and diverse dataset for animal behavior understanding. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19023–19034 (2022) 7
42. Ouyang, W., Dong, Y., Yang, L., Si, J., Pan, X.: I2vedit: First-frame-guided video editing via image-to-video diffusion models. In: SIGGRAPH Asia 2024 Conference Papers. pp. 1–11 (2024) 2, 3
43. Pan, B., Xu, Z., Huang, C.H., Singh, K.K., Zhou, Y., Guibas, L.J., Yang, J.: Actanywhere: Subject-aware video background generation. Advances in Neural Information Processing Systems **37**, 29754–29776 (2024) 4

44. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023) 2
45. Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 724–732 (2016) 7
46. Pitie, F., Kokaram, A.C., Dahyot, R.: N-dimensional probability density function transfer and its application to color transfer. In: ICCV (2005) 3
47. Prolific: Prolific | easily collect high-quality data from real people (2025), https://www.prolific.com/, accessed: 2026-02-15 9
48. Reinhard, E., Adhikhmin, M., Gooch, B., Shirley, P.: Color transfer between images. IEEE Computer graphics and applications **21**(5), 34–41 (2001) 3
49. Ren, W., Yang, H., Zhang, G., Wei, C., Du, X., Huang, W., Chen, W.: Consisti2v: Enhancing visual consistency for image-to-video generation. Transactions on Machine Learning Research (2024), https://openreview.net/forum?id=vqniLmUDvj 2, 3
50. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022) 2, 3
51. Shao, H., Wang, S., Zhou, Y., Song, G., He, D., Zong, Z., Qin, S., Liu, Y., Li, H.: Vividface: A robost and high-fidelity video face swapping framework. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025), https://openreview.net/forum?id=wyv81ezGgv 4
52. Shi, X., Huang, Z., Wang, F.Y., Bian, W., Li, D., Zhang, Y., Zhang, M., Cheung, K.C., See, S., Qin, H., et al.: Motion-i2v: Consistent and controllable image-to-video generation with explicit motion modeling. In: ACM SIGGRAPH 2024 Conference Papers. pp. 1–11 (2024) 2, 3
53. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=St1giarCHLP 2, 3
54. Tu, Y., Luo, H., Chen, X., Ji, S., Bai, X., Zhao, H.: Videoanydoor: High-fidelity video object insertion with precise motion control. In: Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers. pp. 1–11 (2025) 3, 4
55. Wang, J., Zhang, Y., Zou, J., Zeng, Y., Wei, G., Yuan, L., Li, H.: Boximator: Generating rich and controllable motions for video synthesis. In: Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., Berkenkamp, F. (eds.) Proceedings of the 41st International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 235, pp. 52274–52289. PMLR (21–27 Jul 2024), https://proceedings.mlr.press/v235/wang24cr.html 3
56. Wang, J., Sheng, B., Li, P., Jin, Y., Feng, D.D.: Illumination-guided video composition via gradient consistency optimization. Trans. Img. Proc. **28**(10), 5077–5090 (Oct 2019). https://doi.org/10.1109/TIP.2019.2916769, https://doi.org/10.1109/TIP.2019.2916769 4
57. Wang, Y., He, Y., Li, Y., Li, K., Yu, J., Ma, X., Li, X., Chen, G., Chen, X., Wang, Y., Luo, P., Liu, Z., Wang, Y., Wang, L., Qiao, Y.: Internvid: A large-scale video-text dataset for multimodal understanding and generation. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=MLBdiWu4Fw 5, 8

58. Wright, S.: Digital compositing for film and video. Routledge (2013) 2
59. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., ming Yin, S., Bai, S., Xu, X., Chen, Y., Chen, Y., Tang, Z., Zhang, Z., Wang, Z., Yang, A., Yu, B., Cheng, C., Liu, D., Li, D., Zhang, H., Meng, H., Wei, H., Ni, J., Chen, K., Cao, K., Peng, L., Qu, L., Wu, M., Wang, P., Yu, S., Wen, T., Feng, W., Xu, X., Wang, Y., Zhang, Y., Zhu, Y., Wu, Y., Cai, Y., Liu, Z.: Qwen-image technical report (2025), https://arxiv.org/abs/2508.02324 2, 4, 8
60. Wu, J., Li, X., Zeng, Y., Zhang, J., Zhou, Q., Li, Y., Tong, Y., Chen, K.: Motionbooth: Motion-aware customized text-to-video generation. Advances in Neural Information Processing Systems **37**, 34322–34348 (2024) 3
61. Wu, W., Li, Z., Gu, Y., Zhao, R., He, Y., Zhang, D.J., Shou, M.Z., Li, Y., Gao, T., Zhang, D.: Draganything: Motion control for anything using entity representation. In: European Conference on Computer Vision. pp. 331–348. Springer (2024) 3
62. Xu, N., Yang, L., Fan, Y., Yue, D., Liang, Y., Yang, J., Huang, T.: Youtube-vos: A large-scale video object segmentation benchmark. arXiv preprint arXiv:1809.03327 (2018) 7
63. Xue, S., Agarwala, A., Dorsey, J., Rushmeier, H.: Understanding and improving the realism of image composites. ACM Transactions on Graphics **31**(4), 84 (2012) 4
64. Yang, B., Gu, S., Zhang, B., Zhang, T., Chen, X., Sun, X., Chen, D., Wen, F.: Paint by example: Exemplar-based image editing with diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18381–18391 (2023) 4, 8
65. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., Yin, D., Yuxuan.Zhang, Wang, W., Cheng, Y., Xu, B., Gu, X., Dong, Y., Tang, J.: Cogvideox: Text-to-video diffusion models with an expert transformer. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=LQzN6TRFg9 2, 3, 5, 8
66. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3836–3847 (2023) 2
67. Zhang, Y., Wei, Y., Jiang, D., ZHANG, X., Zuo, W., Tian, Q.: Controlvideo: Training-free controllable text-to-video generation. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=5a79AqFr0c 3
68. Zhao, H., Lu, T., Gu, J., Zhang, X., Zheng, Q., Wu, Z., Xu, H., Jiang, Y.G.: Magdiff: Multi-alignment diffusion for high-fidelity video generation and editing. In: European Conference on Computer Vision. pp. 205–221. Springer (2024) 2, 3
69. Zi, B., Peng, W., Qi, X., Wang, J., Zhao, S., Xiao, R., Wong, K.F.: Minimax-remover: Taming bad noise helps video object removal. arXiv preprint arXiv:2505.24873 (2025) 5, 7