LayerT2V: Interactive Multi-Object Trajectory Layering for Video Generation

Kangrui Cen[1] Baixuan Zhao[1] Yi Xin[2,3] Siqi Luo[1]
Guangtao Zhai[1] Xiaohong Liu[1†]

[1]Shanghai Jiao Tong University [2]Nanjing University [3]Shanghai Innovation Institute

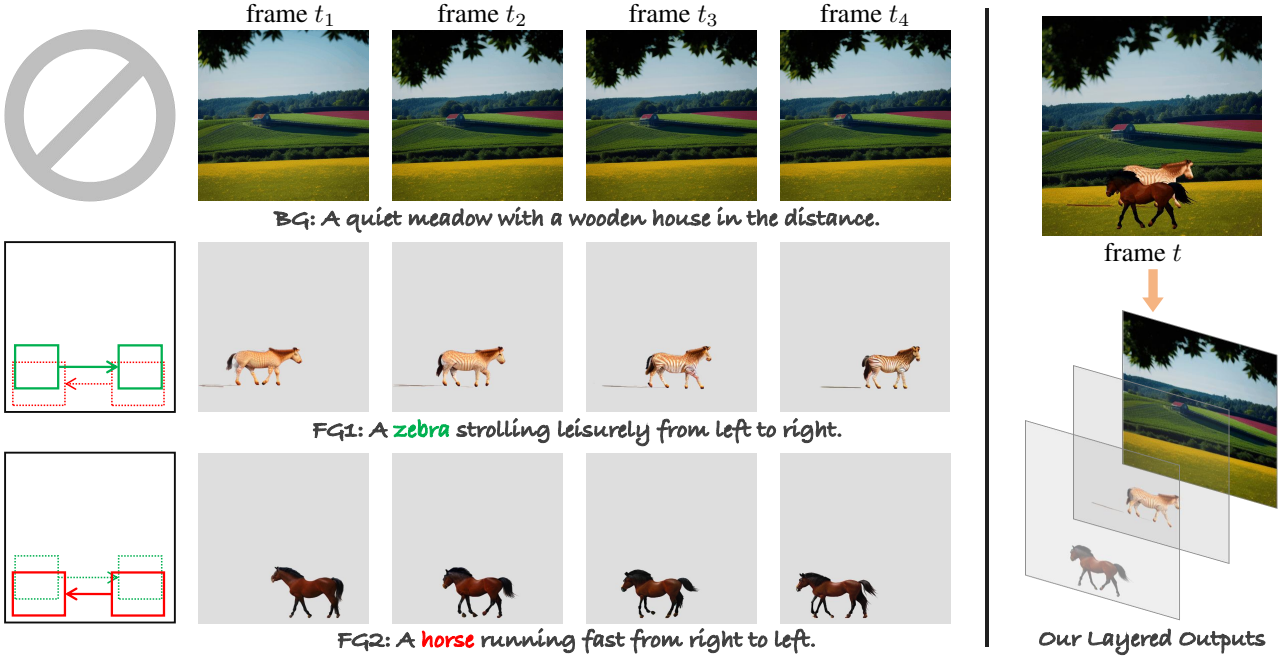


Figure 1. Layered video generation process. Given a long complex prompt and conditional motion trajectories, LayerT2V first exploits the base model to generate the background (BG) video, then generates foreground (FG) subjects layer by layer with our carefully designed Layer-Customized Module to control their motion trajectories. The newly generated layer will be conditioned on all the previous output layers, ensuring excellent harmony and consistency between layers.

Abstract

Controlling object motion trajectories in Text-to-Video (T2V) generation is a challenging and relatively under-explored area, particularly in scenarios involving multiple moving objects. Most community models and datasets in the T2V domain are designed for single-object motion, limiting the performance of current generative models in multi-object tasks. Additionally, existing motion control methods in T2V either lack support for multi-object motion scenes or experience severe performance degradation when object trajectories intersect, primarily due to the semantic conflicts in colliding regions. To address these limitations, we introduce LayerT2V, the first approach for generating video by compositing background and foreground objects layer by layer. This layered generation enables flexible integration of multiple independent elements within a video, positioning each element on a distinct "layer" and thus facilitating coherent multi-object synthesis while enhancing control over the generation process. Extensive experiments demonstrate the superiority of LayerT2V in generating complex multi-object scenarios, showcasing 1.4× and 4.5× improvements in mIoU and AP50 metrics over state-of-the-art (SOTA) methods. Project page and code are available at https://kr-panghu.github.io/LayerT2V/.

1. Introduction

Text-to-Video (T2V) generation has achieved remarkable success in producing diverse, realistic, and high-quality videos through prompt conditioning. A significant factor

†Corresponding author.

contributing to this advancement is the development of Latent Diffusion Models [35], which enable efficient computation in a lower-dimensional latent space and generate high-quality videos by retaining key features while allowing for flexible control over the generation process.

Recently, numerous researchers have focused on customizing generated video content, including camera movements [2, 17, 19, 25], object motions [20, 24, 31, 46, 51, 55] and transformations [48], and identity preservation [13, 22, 44, 49]. Compared to image generation, the core aspect of video generation lies in its intrinsic motion across frames. Thus, controlling object motion trajectories is a critical component in customized video generation. Some training-based methods learn the trajectory of the object through massive object-trajectory pairs [46], or learn a particular type of motion through several reference videos [48, 55]. On the other hand, tuning-free methods usually put efforts into manipulating attention mechanisms [20, 24, 51] or noise resampling [31] to achieve object motion control, utilizing the inherent priors of pretrained T2V models. These methods primarily address the generation of single-object scenes. However, it is worth noting that with the development of customized T2V models, the demand for multi-object control scenarios is steadily increasing. Although existing models demonstrate strong capabilities in controlling object motion, few of them are dedicated to the generation and customization of multi-object scenes, thus lacking support for more complex user-defined control and multi-object motion personalization.

The primary challenge these models encounter in multi-object scenarios is the collision of motion trajectories among objects. This precipitates semantic conflicts in the same region, where a single pixel is conditioned on multiple foreground object prompts. Achieving a solution through training would require a large-scale dataset encompassing complex prompts, multiple object trajectories, and videos. While tuning-free methods also fail to resolve the bottleneck of a pixel being conditioned on multiple foregrounds. A straightforward solution might be to introduce depth information, but rendering a sequence of depth maps that capture the motion of multiple objects is extremely cumbersome and not user-friendly.

Inspired by these issues, we are curious whether it is possible to resolve motion conflicts in multi-object scenes through simple control inputs, such as bounding boxes (bboxes) or point trajectories. To address this, we propose LayerT2V, the first approach that generates video frames in layers. As illustrated in Figure 1, we start by generating the background layer and then progressively generate foreground layers. Each generated layer represents a portion of the prompt with complete semantics. Ultimately, these video layers are stacked together, with the later-generated layers occluding the information from the earlier ones. This approach inherently resolves the issue of motion trajectory conflicts between multiple objects.

Furthermore, we carefully design a Layer-Customized Module, which incorporates the guided cross-attention, oriented attention sharing, and attention isolation, to control the motion trajectory of the generated layer and towards a harmonious blending with the previously generated layers. Moreover, we observe that as the number of foreground layers increases, if the newly generated foreground is fully conditioned on all previously generated layers, it will trigger a redundant consistency issue between multiple foregrounds, leading to integration between foreground objects. To overcome this obstacle, we propose a Harmony-Consistency Bridge, dividing the conditioning process into two stages and resolving the redundant consistency issue between foregrounds.

In summary, our contributions are as follows:

- To the best of our knowledge, LayerT2V is the *first* T2V method from the perspective of video layering and the *first* to address semantic conflicts caused by colliding multi-object motion trajectories.
- To address semantic conflicts in multi-object scenarios, we propose a Layer-Customized Module and a Harmony-Consistency Bridge. These modules collectively facilitate the flexible integration of multiple independent objects, effectively handling multi-object generation while maintaining harmony and consistency across layers.
- We conduct extensive experiments to demonstrate the superiority of LayerT2V over SOTA models both qualitatively and quantitatively, showcasing 1.4-fold and 4.5-fold improvements in the mIoU and AP50 metrics.

2. Related Works

2.1. Text-to-Vision Generation

Text-to-Image (T2I) model enables the automatic synthesis of images from textual descriptions [29, 34, 39] using text embedding model such as CLIP [32]. The performance of T2I models has been significantly enhanced by the introduction of Latent Diffusion Models [35]. This advancement has drawn considerable attention to image customization, with approaches like ControlNet [54] and T2I-Adapter [27] offering enhanced control. Other methods focus on maintaining certain identities [36, 38] or controlling layout in image synthesis [15, 21]. The most recent LayerDiffusion presented by Zhang *et al*. [53] adjusts the latent distribution of pretrained models to support "latent transparency" by training LoRAs [18], showing ability in structured content synthesis like background conditioned generation.

Building on this foundation, Text-to-Video (T2V) generation extends the challenge by requiring not only spatial but also temporal coherence. Several works such as [10, 12, 14, 16, 42, 50] show methods that build on top

Method	Cond Type	Both MT&ST	MultiObj Motion	Colliding Motion
Wang *et al*. [46]	Point	✘	✔	✘
Yin *et al*. [52]	Point	✔	✔	✘
Jain *et al*. [20]	Bbox	✔	✘	✘
Yang *et al*. [51]	Bbox	✔	✔	✘
Ma *et al*. [24]	Bbox	✔	✔	✘
Qiu *et al*. [31]	Bbox	✔	✘	✘
Ours	Bbox	✔	✔	✔

Table 1. Comparison to previous works. Here "MT" and "ST" stand for moving trajectory control and static trajectory control.

of image diffusion models. [4, 23, 40] introduce 3D convolutional layers in the denoising UNet to learn temporal information. Recent T2V models draw inspiration from latent diffusion [35] and operate in a lower-dimensional and more compact latent space to reduce computational complexity. [3] utilizes curated training data and is capable of generating high-quality videos.

2.2. Controllable Motion in Video Generation

As the capabilities of video generation models continue to improve, some outstanding works for controlling object movement have emerged in the community. [48, 55] are trained by employing reference videos for motion customization. However, if we follow the idea of these works for complex multi-object scenarios, it is difficult to obtain an appropriate reference video dataset for training. Plus, it restricts the generated content greatly to the existing videos and leads to insufficient generation freedom. [45] leverages the motion vectors extracted from compressed videos as a direct control signal to impart direction over temporal dynamics, but it needs a compositional depth sequence to control plural objects, which is hugely labor-intensive. [46] encodes the movement paths of objects into a vector field, but it fails to manipulate obstacle trajectories since it directly merges different trajectories into one as a control signal. [20, 24, 31, 47, 51] probe into tuning-free methods by guided attention maps injection, but they failed to control multi-object scenario when the objects collided in a particular area. This is mainly because, in the cross-attention map, the semantic conflict of the coincident region reinforces the attention to different objects at the same time, leading to the deterioration of the generated results.

Different from these models, our LayerT2V focuses on synthesizing one video layer with a single object at one time, then blends them together into the output video to enable multi-object synthesis. In this way, we naturally resolve the conflicts between guided attention map injection and our objective: During the layer generation process, the cross-attention map is modified without distraction because we focus on only one foreground prompt at a time.

3. Methodology

3.1. Preliminaries: Video Diffusion Models

The Latent Diffusion Model (LDM) [35] aims to generate high-quality and diverse images and operate the diffusion process within a latent space to achieve computational efficiency. The most widely used model in the community is Stable Diffusion, which consists of two main components: an image Variational Auto-Encoder (VAE) that converts image representations into and back from a latent space and a denoising UNet that iteratively processes the noisy latent to predict the noise. Some LDM-based video generation methods pass noisy latents through a UNet $\mathbf{U}_\theta$ parametrized by θ to iteratively denoise them, then utilize an image VAE to sequentially decode each latent into an image (i.e., video frame). Denoting the number of iterative denoising steps as T, the process begins with noisy latents $\mathbf{x}_{(T)}$ and progresses to clean latents $\mathbf{x}_{(0)}$ for the VAE decoder. For text-driven generation, the denoising process at time step t can be formulated as

$$\mathbf{x}_{(t-1)} = \mathbf{U}_\theta(\mathbf{x}_{(t)}, \mathbf{c}), \tag{1}$$

where $\mathbf{c}$ is the prompt embeddings.

3.2. LayerT2V

Overall pipeline. Our overall pipeline is illustrated in Figure 2 where we perform the foreground generation above a pre-generated background. The background video layer is initially processed through control convolutions and then integrated into latents to support background-to-foreground conditioned generation. For foreground motion control, users specify bbox sequences with foreground prompts which are processed by Layer-Customized Module to ensure precise motion control and seamless blending. After the denoising process, latents are decoded by a transparent decoder to produce a sequence of transparent foreground frames. In cases where additional foreground elements are present, the Harmony-Consistency Bridge manipulates control signals to enable a two-stage conditioned generation approach. Ultimately, all generated foregrounds are layered onto the background and passed through a harmonizer for texture refinement, rendering a cohesive multi-object scene with versatile motion control.

3.2.1. Layered Video Synthesis

As mentioned in 3.1, the VAE in the image generation process can be reused in the video domain, and [53] trains transparent LoRAs that adjust latent distribution to support alpha channel and an additional VAE to support transparent image encoding/decoding. We adopt the pre-trained modules in [53] with its LoRAs and decoder $\mathcal{D}$ besides the original decoder $\mathcal{D}^*$ and inflate it into video generation model to enable transparent video synthesis. Suppose $\mathbf{x}$ is the latents denoised by a pretrained UNet. By fine-tuning with

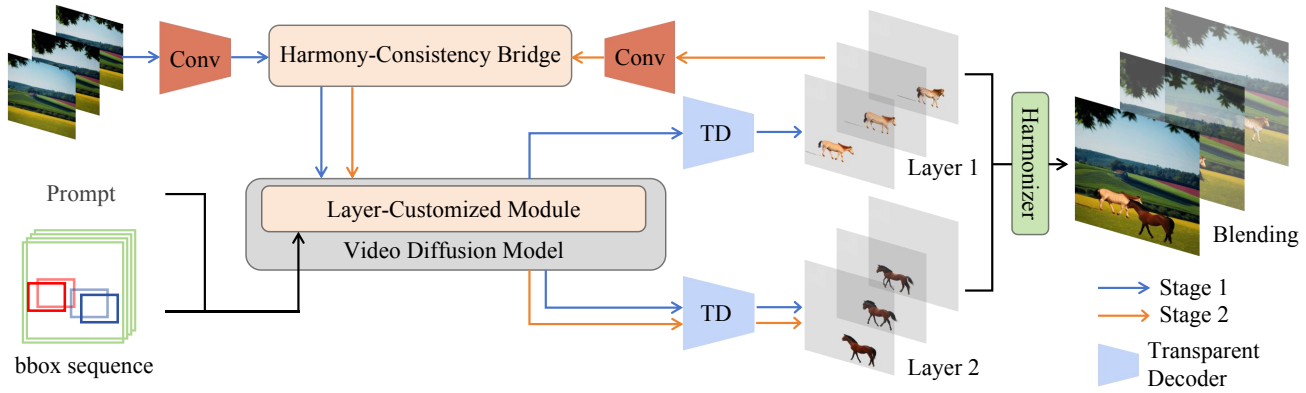


Figure 2. The overall structure of the LayerT2V. This pipeline supports both background video generation and multiple foreground video layers generation. We propose a Layer-Customized Module to control the motion trajectories of the subjects with background conditioned generation. The proposed Harmony-Consistency Bridge coordinates consistency between Layer 2 and Layer 1. In stage one, only background frames are input into control convolutions, while in stage two, both background frames and Layer 1 are incorporated.

a dataset of transparent images, denoising UNet can obtain an adjusted latent $\mathbf{x}_a = \mathbf{x} + \mathbf{x}_\epsilon$ with an offset $\mathbf{x}_\epsilon$ in a modified latent space for transparent frames. The decoded RGB reconstruction can be denoted as $\hat{I} = \mathcal{D}^*(\mathbf{x}_a)$, thus the reconstruction of RGBA frames can be denoted as $[\hat{I}_c, \hat{I}_a] = \mathcal{D}(\hat{I}, \mathbf{x}_a)$, where $\hat{I}_c$ and $\hat{I}_a$ represent the RGB and alpha values respectively.

To enable background conditioned generation, we first utilize the base T2V model to generate background video $\mathbf{b}$ with f frames and input noisy latents $\mathbf{x} = [\mathbf{x}^{FG}, \mathbf{x}^{BG}]$ with shape $[f, 2, h, w, ch]$ which represents a combination of foreground latents $\mathbf{x}^{FG}$ and background latents $\mathbf{x}^{BG}$ in denoising process. Here h, w, ch represent the height, width, and number of channels, respectively. Then, we use control convolutions to embed $\mathbf{b}$ into $\mathbf{x}^{BG}$. To formulate, our background conditioned generation is based on the following transforms

$$\mathbf{x}^{BG} \leftarrow \mathbf{x}^{BG} \oplus \text{Conv}(\mathbf{b}). \tag{2}$$

Given the bbox sequence $\mathcal{B} = [B_1, B_2, \ldots, B_f]$, we have

$$\mathbf{x}_{(t-1)} = \mathbf{U}_{\theta'}(\mathbf{x}_{(t)}, \mathbf{c}, \mathbf{b}, \mathcal{B}). \tag{3}$$

Here $\mathbf{U}_{\theta'}$ represents the backbone video diffusion UNet incorporated with transparent LoRAs.

3.2.2. Layer-Customized Module

We carefully designed our Layer-Customized Module (LCM) to achieve our goal of handling multi-object scenario generation while preserving harmony and consistency between multiple layers. LCM has two vital components, guided cross-attention and oriented temporal-attention. In guided cross-attention, we use an attention map injection method with key-frame amplification to align the motion of the output transparent layer with the given bbox sequence. In oriented temporal attention, we separate it into attention-sharing with masks at frame-pixel-wise level and attention-isolation. The former is to enable foreground (FG) latents to maintain consistency with background (BG) latents, like rendering illumination, shadow effects, or color harmony. The latter is to avoid transparent latent distribution, which was offered by transparent LoRA, being disrupted.

Guided Spatial Cross-Attention. Spatial cross-attention plays an essential role in T2V generation, as it serves as the only pathway for the prompt to embed into the latent representations. Therefore, it is crucial to thoroughly investigate methods to steer spatial cross-attention towards our desired outcomes. Rather than using a linear additive mask, we employ a Gaussian function to smoothly construct an additive mask corresponding to the bbox area with an influence coefficient λ, allowing it to guide without negatively disrupting the attention values, thereby preserving the quality of generated content.

In addition, for complex trajectories, simple guidance may not suffice to achieve precise bbox–prompt alignment. To address this, we introduce the concept of key-frames. For example, in a polyline trajectory, the start point, end point, and turning point determine the overall movement of the object and are critical to bbox–prompt alignment. Consequently, we amplify the additive mask of these key-frames by a reinforcement parameter γ_{key} to ensure more effective guidance. For each frame $f_0 \in [f]$, we have a bbox $B \in \mathcal{B}$ and foreground prompt embeddings $\mathbf{c}_F$. Given query Q derived from visual tokens, key K and value V mapped from text embeddings, the guided cross-attention is formulated as

$$\text{CrossAttn}(Q, K, V) = (\text{Softmax}(\mathbf{A}(Q, K)) + \lambda\mathcal{M})V. \tag{4}$$

Here

$$\mathbf{A}[i, j] = \begin{cases} QK^T/\sqrt{d} - \infty, & \text{if } (i, j) \in \Phi(1), \\ QK^T/\sqrt{d}, & \text{otherwise}, \end{cases} \tag{5}$$

where $\Phi(n) = \{(i, j) | \mathbb{I}[i \in B] + \mathbb{I}[j \in \mathbf{c}_F] = n\}$, $\mathbb{I}$ is a

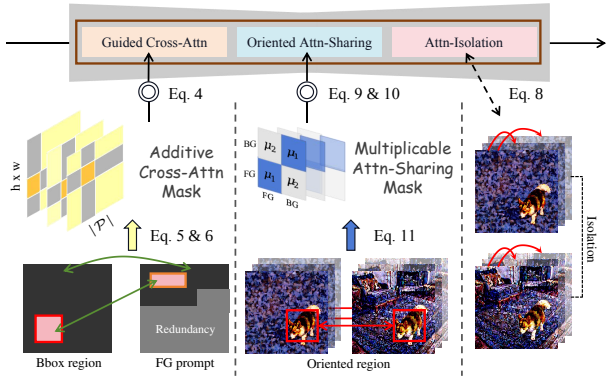


Figure 3. Layer-Customized Module with its three major components: Guided Cross-Attn, Oriented Attn-Sharing, and Attn-Isolation.

boolean indicator, and

$$\mathcal{M}[i,j] = \begin{cases} g_B(i,j) \cdot \gamma(f_0), & \text{if } (i,j) \in \Phi(2), \\ 0, & \text{otherwise}, \end{cases} \tag{6}$$

where g_B is a Gaussian weight [24] within bbox B, $\gamma(f_0) = \gamma_{key} \geq 1$ if f_0 is a key-frame, otherwise $\gamma(f_0) = 1$.

Attention-Sharing and Attention-Isolation. Similar to most UNet-based video diffusion models equipped with spatial attention and temporal attention, our network basically adopts the same methodology but divides the temporal module into two parts: Attention-Sharing and Attention-Isolation. The latents $\mathbf{x}$ will be processed by these sequences of spatio-temporal modules. For easier illustration, we only list the core components, Temporal Transformer (TT) and Spatial Transformer (ST), in the pipeline to formularize the denoising process at time step t as

$$\mathbf{x}_{(t-1)} \leftarrow \text{TT}_{\text{AI}}(\text{TT}_{\text{AS}}(\text{ST}(\mathbf{x}_{(t)}, \mathbf{c}))), \tag{7}$$

where the subscripts AI and AS denote attention-isolation and attention-sharing, and $\mathbf{c}$ indicates the prompt embeddings. Furthermore, we design our TT_{AI} and TT_{AS} as:

$$\text{TT}_{\text{AI}} = \begin{cases} \text{SelfAttn}(\mathbf{x}^{FG}_{[f]}), \\ \text{SelfAttn}(\mathbf{x}^{BG}_{[f]}). \end{cases} \tag{8}$$

$$\text{TT}_{\text{AS}} = \text{SelfAttn}([\mathbf{x}^{FG}_i, \mathbf{x}^{BG}_i])_{i \in [f]}, \tag{9}$$

where $\mathbf{x}^{FG}_{[f]}, \mathbf{x}^{BG}_{[f]}$ denote the separate concatenation of FG and BG latents respectively, and $[\mathbf{x}^{FG}_i, \mathbf{x}^{BG}_i]$ is the combined concatenation of FG and BG latents for i-th frame. As shown in Figure 3, we aggregate all FG and BG latents for frame-pixel-wise attention sharing and then isolate them to perform cross-frame attention.

Oriented Attention-Sharing. In our layer generation framework, the fusion of foreground and background is a critical focus. We aim to avoid a scenario where the foreground layer appears to "float" above the background layer without interactions, thereby losing visual concordance. To address this, we introduce an orientation step within the attention-sharing process, guiding foreground pixels inside the bbox to attend more closely to their corresponding background pixels as illustrated in Figure 3. Formally, for frame f_0 and its bbox B, given query Q, key K, value V in temporal attention,

$$\text{AttnSharing}(Q,K,V) = (\text{Softmax}(\frac{QK^T}{\sqrt{d}}) \odot \mathcal{W}(B))V \tag{10}$$

where $\odot$ denotes the Hadamard (element-wise) product that scales the foreground and background elements in attention-sharing maps: For each frame f_0, and each latent pixel (x,y), the corresponding *frame-pixel-wise* element $\mathcal{W}(B)$ can be formulated as

$$\mathcal{W}(B)[i,j] = \begin{cases} \mu_1 & \text{if } (x,y) \in B, i = j, \\ \mu_2 & \text{if } (x,y) \in B, i \neq j, \\ 1 & \text{otherwise.} \end{cases} \tag{11}$$

Here $\mu_1, \mu_2 \geq 1$ applied to cases "$i = j$" and "$i \neq j$" are two coefficients to strengthen the attention inside bbox area, and the influence of the interaction of foreground and background respectively.

3.2.3. Harmony-Consistency Bridge

We discover that due to the training process of transparent module [53] is demonstrated on {*text, foreground layer, background layer*} pairs with a single object in the foreground layer, the attention sharing will have an extremely negative impact on the newly generated layer if its trajectory is collided with previously generated foreground layer.

Assuming we have a background layer (BG) and the first foreground layer (FG1), we blend them and feed this combined background signal into our pipeline to generate the second foreground layer (FG2), where FG1 and FG2 exhibit substantial motion collision. Under this circumstance, FG2 tends to focus on the motion and texture of FG1 ascribed to attention-sharing, potentially leading FG2 to adopt a similar pattern to FG1. Also, the collision between different foregrounds can disrupt the transparent latent distribution. Based on this observation, we propose a Harmony-Consistency Bridge (HCB) to tackle the temporal redundant consistency between different foregrounds. Since the early steps of the denoising process generate coarse layouts that critically influence the object's motion [49], HCB first conditions the denoising solely on the background layer to ensure accurate motion information. Then, it conditions all previously generated layers to ensure seamless integration of the new layer with the existing content. In summary, there will be two stages in the HCB. Suppose we have obtained BG layer and $(i-1)$ FG layers, denote timestep as

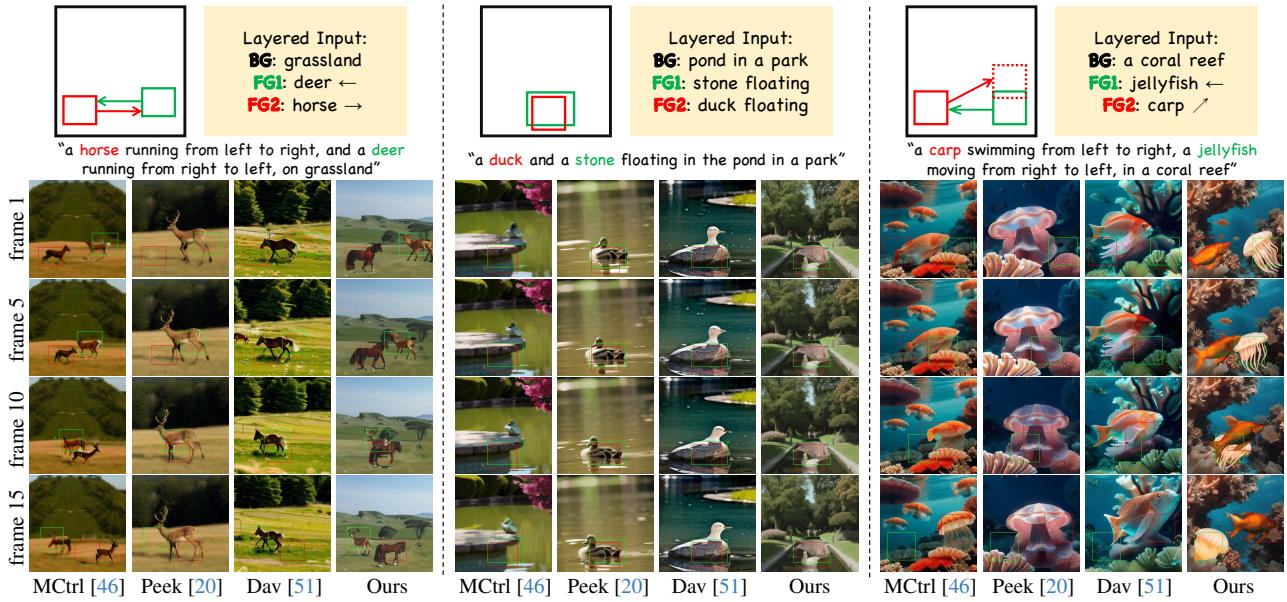


Figure 4. Qualitative comparison on colliding object motion control with current SOTAs. Our method excels in handling cases involving more than one object with colliding motion. Furthermore, we take both static trajectory and moving trajectory into account and showcase that LayerT2V outperforms others.

Methods	Quality		Semantic Fidelity		Trajectory Control			
	FID(↓)	FVD(↓)	CLIPSIM(↑)	Preference(↑)	mIoU(↑)	AP50(↑)	Cov(↑)	CD(↓)
MotionCtrl [46]	153.27	1516.27	30.18	2.22%	6.97	3.04	0.83	0.16
Peekaboo [20]	147.49	1436.12	30.45	2.78%	10.43	0.97	0.84	0.14
Direct-a-Video [51]	140.14	1380.79	29.19	6.67%	12.64	2.05	0.75	0.13
LayerT2V(Ours)	**136.12**	**1356.38**	**32.47**	**88.3%**	**30.12**	**16.62**	**1.00**	**0.05**

Table 2. Quantitative comparison on colliding object motion control with related baselines. Our LayerT2V outperforms competing approaches in multi-object motion scenes, achieving impressive improvements of 1.4× in mIoU and 4.5× in AP50, while maintaining 100% semantic integrity without compromising video quality. [**Bold text**: Best; **Bold red text**: Best and significantly surpass others]

t, and t_ε determines the timestep to transfer to the second stage. For the i-th FG layer, during the first stage ($t > t_\varepsilon$),

$$\mathbf{x}_{(t-1)} = \mathbf{U}_{\theta'}(\mathbf{x}_{(t)}, \mathbf{c}, \mathbf{b}), \tag{12}$$

and for the second stage ($t \leq t_\varepsilon$),

$$\mathbf{x}_{(t-1)} = \mathbf{U}_{\theta'}(\mathbf{x}_{(t)}, \mathbf{c}, \mathbf{b} \circ \mathbf{fg}_1 \circ \cdots \circ \mathbf{fg}_{i-1}). \tag{13}$$

Here $\circ$ denotes blending operation for layers and $\mathbf{b} \circ \mathbf{fg}_1 \circ \cdots \circ \mathbf{fg}_{i-1}$ is the blending of previously generated layers.

4. Experiments

In this section, we present both qualitative (Sec.4.2) and quantitative (Sec.4.3) analysis for LayerT2V, mainly concentrating on the scenario in which the motion trajectories of multiple objects are colliding.

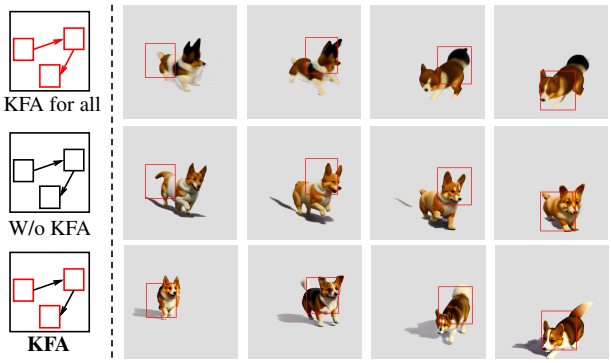
4.1. Settings

Based on our investigation in Table 1, we compare our LayerT2V to three leading methods: MotionCtrl [46], Direct-a-Video [51] and Peekaboo [20]. Peekaboo manipulates spatial-temporal attention outputs and is originally implemented to control a single object. We directly sum multiple bbox sequences into one as the input mask for Peekaboo. Pursuing a justified comparison, we implement these methods all in SDv1.5 UNet-2D backbone inflated with the transformer temporal module of AnimateDiff [11]. After obtaining the transparent frames, we can easily obtain the foreground mask by selecting regions in the alpha map where the alpha values are close to the maximum value of 255. Then we pass the blended frames through INR-Harmonization [8], which enables masked pixel-to-pixel harmonization via continuous image representation. We will provide further explanations in appendix.

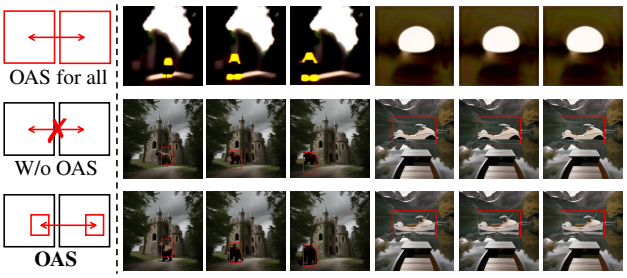
For object motion control, we curate 20 trajectory combos with two to three trajectories each and generate samples using these combos to perform qualitative and quantitative evaluations. For each trajectory combo, we evaluate 10 to 12 different {*background prompt, foreground prompts*} layered inputs with reasonable semantics between layers.

4.2. Qualitative Analysis

The visualized comparison in Figure 4 reveals that LayerT2V can generate multiple objects whose movements are both static and dynamic, or colliding both partially and completely. Without compromising video quality and har-



(a) We aim to control the "corgi" to first move forward, then reverse direction. While both the proposed method and applying KFA across all frames accomplish this, the latter substantially compromises foreground quality.



(b) Uniformly applying OAS disrupts the latent distribution. In contrast, the proposed method improves motion control and harmony. In these examples, the bear darkens naturally under tree shade, and the boat's reflection aligns remarkably well with the lake, enhancing content realism.

Figure 5. Visualized effects of Key-Frame Amplification (KFA) and Oriented Attention-Sharing (OAS).

mony between different layers, our models are capable of blending the foreground seamlessly into the background. For instance, in the 8th column in Figure 4, where we first generate a pond in a park, and then we discover that the subsequently generated stone and duck have some inverted reflection and fit in well with the pond.

Additionally, when two objects collide, baseline models tend to encounter two challenges: semantic mixing and absence [51]. Semantic mixing refers to the incorrect assignment of prompt attributes to different objects. This is exemplified in the 7th column, where the texture of the duck mixes with the stone, and the 9th column, where the carp in the first few frames ultimately transforms into a jellyfish. Semantic absence, a known limitation in T2V models [7], arises when an object fails to appear as expected. For instance, in the 2nd, 3rd, and 10th columns, only one main object appears and the other foreground is missing, leading to huge discordance with prompts. In contrast, our method effectively mitigates these issues, enabling improved control over the motion of multiple foreground objects.

Methods	FID(↓)	CLIPSIM(↑)	mIoU(↑)	CD(↓)
KFA for all	162.28	30.34	27.30	0.11
W/o KFA	140.70	30.27	26.65	0.12
KFA	**136.12**	**32.47**	**30.12**	**0.05**

(a) We evaluate the effectiveness of key-frame amplification in three aspects: frame quality, semantic similarity and bbox-object alignment.

Methods	Preference(↑)	CLIPSIM(↑)	mIoU(↑)	CD(↓)
W/o OAS	5.3%	31.71	27.54	0.08
OAS	**94.7%**	**32.47**	**30.12**	**0.05**

(b) Compared to the quality of the video itself, OSA has a greater impact on the aesthetic value created by the integration of the foreground layer with the background. Thus we conduct a user survey with 15 volunteers to calculate user preference.

Table 3. Quantitative evaluations for Layer-Customized Module.

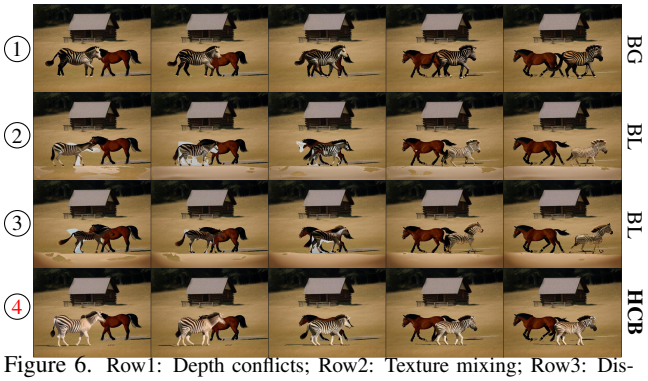


Figure 6. Row1: Depth conflicts; Row2: Texture mixing; Row3: Disruption of transparent latent distribution; Row4: Harmonious multi-object scenes (Ours).

Methods	FID(↓)	FVD(↓)	CLIPSIM(↑)	mIoU(↑)	CD(↓)
Solely BG	138.94	1434.36	32.12	29.28	0.05
Solely BL	141.31	1619.51	28.55	21.90	0.10
HCB	**136.12**	**1356.38**	**32.47**	**30.12**	**0.05**

Table 4. Quantitative evaluations for HCB. Here BL stands for blended conditional information, where we integrate all of the previously generated layers into control convolutions.

4.3. Quantitative Analysis

Evaluation Metrics. We thoroughly perform the evaluation in three aspects: video quality, semantic fidelity, and trajectory control. **(1)** The quality of the generated videos is evaluated using Fréchet Inception Distance (FID) [37] and Fréchet Video Distance (FVD) [41], which are two commonly used metrics for video quality assessment, against the random selected 800 videos in AnimalKingdom [28] dataset. **(2)** We calculate CLIP Similarity (CLIPSIM) [33] between overall prompts and video frames, which measures the semantic similarity between the input prompt and the generated video. Besides, we conduct a user study involving 15 participants, where each sample set consists of four videos generated under the same settings by four methods. Participants were asked to select the best result based on three dimension: (i) Semantic integrity, (ii) semantic clarity, and (iii) alignment with the prompt. **(3)** We utilize OWL-ViT-large [26] to evaluate the motion trajectory control and corresponding four metrics following the methodology in Jain *et al.* [20]: Coverage (Cov), mean of Intersection-

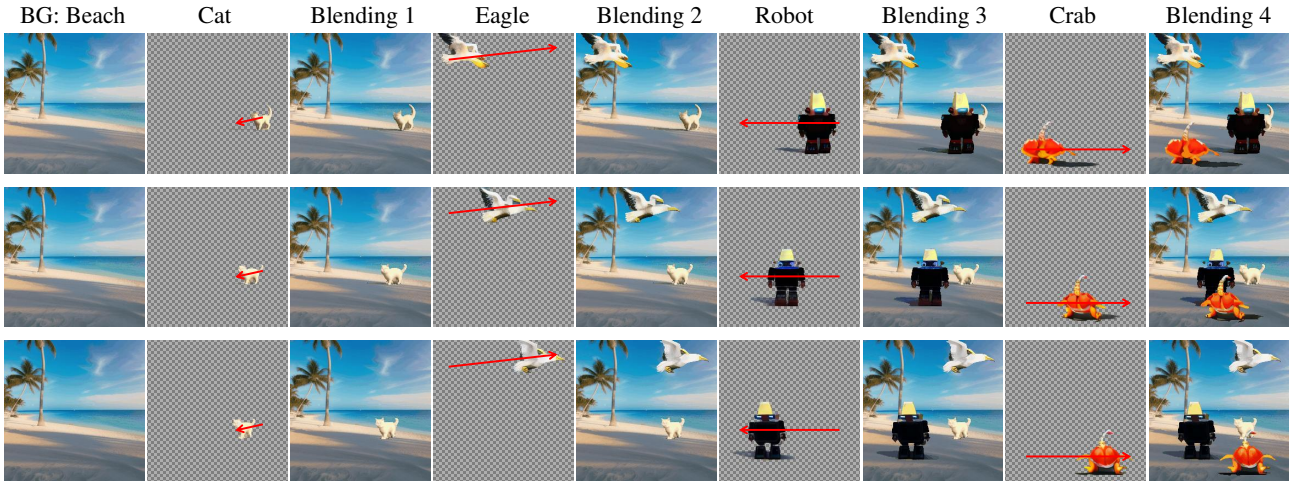


Figure 7. More iterations. We present a case to show the ability of LayerT2V to handle multiple object generation iteratively. For convenience, we use arrows to replace bounding boxes.

over-Union (mIoU) against the input boxes, Centroid Distance (CD) and average precision at 50% IoU threshold (AP50). Here, Cov and CD represent the fraction of generated videos that the bboxes detected in more than half of the frames and the distance between the centroid of the generated subject and input mask, respectively.

Colliding Motion Control. We assess colliding motion for multi-object scenarios and report the evaluation results in Table 2. Baseline models generally display two main issues when handling multi-object colliding motion control: (1) the generated output deviates significantly from the prompt, or only a subset of the subjects specified in the prompt are rendered, leading to substantial declines in CLIPSIM and user preference; (2) ineffective trajectory control, where subjects frequently fail to adhere to the prescribed motion paths. Our model overcomes these limitations impressively, adhering strictly to the semantics and trajectories of conditioned information. Notably, our model outperforms all baselines across *all Semantic Fidelity and Trajectory Control metrics*. Without compromising video quality, our model slightly surpasses the others in FID and FVD, showing strong capability for colliding motion control.

4.4. Ablation Study

With respect to the trajectory customization part of LCM, we find out that the absence of key-frame amplification or simply applying key-frame amplification to all frames will cause the degradation of object localization and the correctness of alpha values. Visualized results can be found in Figure 5a and the quantitative analysis is present in Table 3a. Similarly, Figure 5b suggests the ability of oriented attention-sharing to achieve harmonious blending between layers. Since shadow, reflection, and illuminating are more about the aesthetic values of a video, we conduct user studies within 15 volunteers and collect their preference based on: 1) layer quality; 2) bbox alignment; 3) blending harmonization, and report the average preference in Table 3b. Furthermore, other than our proposed HCB module, we simply feed the pure background layer or the blending of all the previous output layers as the control signal. In Figure 6, we fix the background and the first foreground while adjusting the methods to synthesize the second foreground. We observe that solely using the background layer can lead to depth conflicts between foregrounds, while using only the blending may disrupt transparent latent distribution or cause texture mixing. Quantitative results are reported in Table 4.

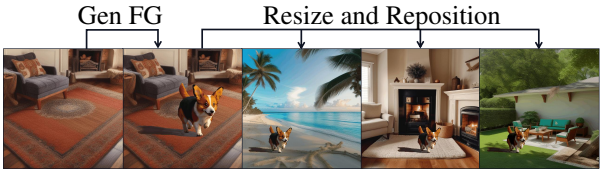


Figure 8. Layer transplantation demonstrates that our model has a wide range of promising application scenarios.

Extensive Blending. The inherent capability of layer generation enables the iterative creation of additional video layers. By strategically providing appropriate bbox sequences and prompts, informed by the semantic context of the generated background, we can generate complex multi-object motion patterns, as demonstrated in Figure 7. This represents a significant advancement over traditional T2V models, which are limited in controlling multiple objects.

Layer Transplantation. We have observed that generated layers, even when lacking specialized attributes like reflections, can offer significant practical value for transplantation to other videos. Moreover, the transparency of these generated layers allows for flexible scaling, repositioning, and seamless overlay onto diverse backgrounds. This versatility, akin to techniques employed in video editing, is illustrated in Figure 8.

Interactions between foregrounds. Our primary goal is to

address colliding motion control in T2V. Each foreground is generated independently if they exist at different depths. If interactions within the same depth are desired, they can be achieved by grouping multiple objects together and generating these foregrounds simultaneously. This approach can be seen as an extension of LayerT2V and is illustrated in Fig. 9. For example, in the first line, initially, the ball occludes the dog (highlighted in yellow); subsequently, the dog occludes the ball (highlighted in blue). However, this method presents instability, which will be discussed in Sec. 5. Semantic conflicts in overlapping regions intensify attention competition among different objects, sometimes leading to failures in generating all intended objects.

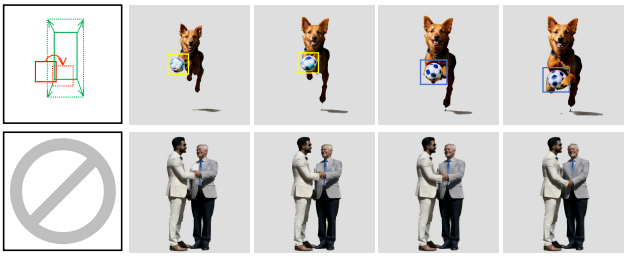


Figure 9. Interaction within the same depth: (1) "A dog is chasing a ball", these two FGs are combined into a group and synthesized together; (2) "**Two gentlemen** shake hands", these two FGs are combined and generated together without explicit bbox prompt.

5. Limitations and Future Work

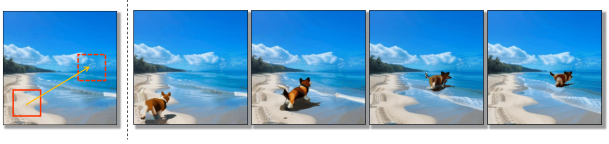


Figure 10. Limitations caused by semantic conflicts between bbox and background.

The foreground generation of LayerT2V depends on background features, so mismatches between bboxes and background semantics can degrade quality and produce unrealistic outputs. This can be mitigated by providing control conditions that align more closely with the background. For instance, in Figure 10, our bbox sequence extends from the beach to the distant sea and sky, which significantly violates the semantic consistency of the background. This situation will result in an unnatural appearance.

Moreover, it is worth noting that the experimental T2V backbone employed in our work is somewhat outdated, which significantly constrains the visual quality and overall capability of our model. As part of our future work, we plan to implement the proposed approach on more recent DiT-based models. This upgrade is expected to enhance resolution, improve video quality, reduce artifacts and hallucinations, and achieve better motion consistency.

6. Conclusion

This paper presents LayerT2V, the first T2V model adopts the methodology of video layering. By overlaying coherent transparent video layers onto a pre-generated background, LayerT2V addresses the challenge of controlling multi-object motion trajectories, especially in handling colliding motions. Experimental evaluations show 1.4-fold and 4.5-fold improvements in the mIoU and AP50 metrics for motion control over current SOTA and significant gains in other metrics. In summary, LayerT2V provides a novel solution for generating complex multi-object interaction scenes.

Acknowledgements. We would like to greatly thank Ming-Hsuan Yang and Kelvin C.K. Chan for their insightful discussions and generous support. We also thank Yangnan Lin for his help in testing benchmarks for our model.

References

[1] Aishwarya Agarwal, Srikrishna Karanam, K J Joseph, Apoorv Saxena, Koustava Goswami, and Balaji Vasan Srinivasan. A-star: Test-time attention segregation and retention for text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2283–2293, 2023. 1

[2] Sherwin Bahmani, Ivan Skorokhodov, Aliaksandr Siarohin, Willi Menapace, Guocheng Qian, Michael Vasilkovsky, Hsin-Ying Lee, Chaoyang Wang, Jiaxu Zou, Andrea Tagliasacchi, et al. Vd3d: Taming large video diffusion transformers for 3d camera control. *arXiv preprint arXiv:2407.12781*, 2024. 2

[3] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 3

[4] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023. 3

[5] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22560–22570, 2023. 1

[6] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Trans. Graph.*, 42(4), 2023. 3, 5

[7] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–10, 2023. 7

[8] Jianqi Chen, Yilan Zhang, Zhengxia Zou, Keyan Chen, and Zhenwei Shi. Dense pixel-to-pixel harmonization via continuous image representation. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(5):3876–3890, 2024. 6, 1

[9] Dave Epstein, Allan Jabri, Ben Poole, Alexei Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation. In *Advances in Neural Information Processing Systems*, pages 16222–16239. Curran Associates, Inc., 2023. 1

[10] Songwei Ge, Seungjun Nah, Guilin Liu, Tyler Poon, Andrew Tao, Bryan Catanzaro, David Jacobs, Jia-Bin Huang, Ming-Yu Liu, and Yogesh Balaji. Preserve your own correlation: A noise prior for video diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22930–22941, 2023. 2

[11] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 6, 1, 5

[12] William Harvey, Saeid Naderiparizi, Vaden Masrani, Christian Weilbach, and Frank Wood. Flexible diffusion modeling of long videos. *Advances in Neural Information Processing Systems*, 35:27953–27965, 2022. 2

[13] Xuanhua He, Quande Liu, Shengju Qian, Xin Wang, Tao Hu, Ke Cao, Keyu Yan, and Jie Zhang. Id-animator: Zero-shot identity-preserving human video generation. *arXiv preprint arXiv:2404.15275*, 2024. 2

[14] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022. 2

[15] Jiun Tian Hoe, Xudong Jiang, Chee Seng Chan, Yap-Peng Tan, and Weipeng Hu. Interactdiffusion: Interaction control in text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6180–6189, 2024. 2

[16] Tobias Höppe, Arash Mehrjou, Stefan Bauer, Didrik Nielsen, and Andrea Dittadi. Diffusion models for video prediction and infilling. *arXiv preprint arXiv:2206.07696*, 2022. 2

[17] Chen Hou, Guoqiang Wei, Yan Zeng, and Zhibo Chen. Training-free camera control for video generation. *arXiv preprint arXiv:2406.10126*, 2024. 2

[18] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 2

[19] Teng Hu, Jiangning Zhang, Ran Yi, Yating Wang, Hongrui Huang, Jieyu Weng, Yabiao Wang, and Lizhuang Ma. Motionmaster: Training-free camera motion transfer for video generation. *arXiv preprint arXiv:2404.15789*, 2024. 2

[20] Yash Jain, Anshul Nasery, Vibhav Vineet, and Harkirat Behl. Peekaboo: Interactive video generation via masked-diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8079–8088, 2024. 2, 3, 6, 7, 1

[21] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521, 2023. 2

[22] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8640–8650, 2024. 2

[23] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. Videofusion: Decomposed diffusion models for high-quality video generation. *arXiv preprint arXiv:2303.08320*, 2023. 3

[24] Wan-Duo Kurt Ma, John P Lewis, and W Bastiaan Kleijn. Trailblazer: Trajectory control for diffusion-based video generation. *arXiv preprint arXiv:2401.00896*, 2023. 2, 3, 5, 1

[25] Andrew Marmon, Grant Schindler, José Lezama, Dan Kondratyuk, Bryan Seybold, and Irfan Essa. Camvig: Camera aware image-to-video generation with multimodal transformers. *arXiv preprint arXiv:2405.13195*, 2024. 2

[26] M Minderer, A Gritsenko, A Stone, M Neumann, D Weissenborn, A Dosovitskiy, A Mahendran, A Arnab, M Dehghani, Z Shen, et al. Simple open-vocabulary object detection with vision transformers. arxiv 2022. *arXiv preprint arXiv:2205.06230*, 2, 2022. 7

[27] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4296–4304, 2024. 2

[28] 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. 7, 1

[29] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2

[30] Quynh Phung, Songwei Ge, and Jia-Bin Huang. Grounded text-to-image synthesis with attention refocusing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7932–7942, 2024. 1

[31] Haonan Qiu, Zhaoxi Chen, Zhouxia Wang, Yingqing He, Menghan Xia, and Ziwei Liu. Freetraj: Tuning-free trajectory control in video diffusion models. *arXiv preprint arXiv:2406.16863*, 2024. 2, 3

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry,

Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 7, 1

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

[35] 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. 2, 3

[36] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 2

[37] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 7, 1

[38] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 2

[39] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 2

[40] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 3

[41] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 7, 1

[42] Vikram Voleti, Alexia Jolicoeur-Martineau, and Chris Pal. Mcvd-masked conditional video diffusion for prediction, generation, and interpolation. *Advances in neural information processing systems*, 35:23371–23385, 2022. 2

[43] Jiawei Wang, Yuchen Zhang, Jiaxin Zou, Yan Zeng, Guoqiang Wei, Liping Yuan, and Hang Li. Boximator: Generating rich and controllable motions for video synthesis. *arXiv preprint arXiv:2402.01566*, 2024. 1

[44] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, Anthony Chen, Huaxia Li, Xu Tang, and Yao Hu. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024. 2

[45] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[46] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 2, 3, 6, 1

[47] 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. *arXiv preprint arXiv:2406.17758*, 2024. 3

[48] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7623–7633, 2023. 2, 3

[49] Tao Wu, Yong Zhang, Xintao Wang, Xianpan Zhou, Guangcong Zheng, Zhongang Qi, Ying Shan, and Xi Li. Customcrafter: Customized video generation with preserving motion and concept composition abilities. *arXiv preprint arXiv:2408.13239*, 2024. 2, 5

[50] Ruihan Yang, Prakhar Srivastava, and Stephan Mandt. Diffusion probabilistic modeling for video generation. *Entropy*, 25(10):1469, 2023. 2

[51] Shiyuan Yang, Liang Hou, Haibin Huang, Chongyang Ma, Pengfei Wan, Di Zhang, Xiaodong Chen, and Jing Liao. Direct-a-video: Customized video generation with user-directed camera movement and object motion. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–12, 2024. 2, 3, 6, 7, 1

[52] Shengming Yin, Chenfei Wu, Jian Liang, Jie Shi, Houqiang Li, Gong Ming, and Nan Duan. Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. *arXiv preprint arXiv:2308.08089*, 2023. 3

[53] Lvmin Zhang and Maneesh Agrawala. Transparent image layer diffusion using latent transparency. *arXiv preprint arXiv:2402.17113*, 2024. 2, 3, 5

[54] 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. 2

[55] Rui Zhao, Yuchao Gu, Jay Zhangjie Wu, David Junhao Zhang, Jia-Wei Liu, Weijia Wu, Jussi Keppo, and Mike Zheng Shou. Motiondirector: Motion customization of text-to-video diffusion models. In *European Conference on Computer Vision*, pages 273–290. Springer, 2025. 2, 3

LayerT2V: Interactive Multi-Object Trajectory Layering for Video Generation

Supplementary Material

A. Implementation Details

Control signal input for different methods. A bbox sequence is primarily determined by several key-frames, with intermediate bboxes obtained through interpolation. This bbox sequence is then scaled to match the corresponding latent space dimensions (h', w') and processed using attention weight injection (for LayerT2V and Direct-a-Video [51]) or masking (for Peekaboo [20]). For MotionCtrl [46], we calculate the centers of the bbox sequence to obtain a point sequence and use the provided scripts to generate motion vectors as conditions. For a fair comparison, all methods are implemented on SDv1.5 inflated by AnimateDiff [11], which introduces temporal transformer modules to the text-to-image backbone.

Masked harmonization. After generating the transparent frames, the alpha values (α) at foreground pixel locations are typically high (generally close to maximum 255), while non-foreground regions may still retain small alpha values rather than being fully transparent (e.g., $\alpha = 10$). Therefore, we first apply a threshold-based filtering process to these frames, resetting regions with very small alpha values to $\alpha = 0$. This allows us to easily obtain the foreground region mask from the alpha map. Next, we input the blended frames and foreground masks into INR-Harmonization [8], which enhances the alignment between the foreground and background through pixel-to-pixel processing and improves texture realism, as shown in Fig. 13.

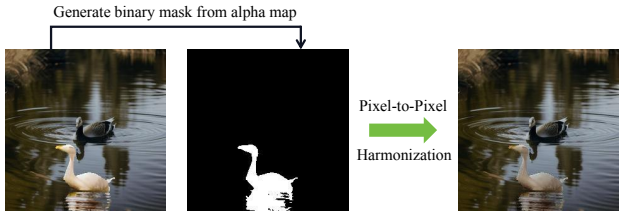


Figure 13. Illustration of masked harmonization.

The generated videos for evaluation have a resolution of 256×256 and a length of 16 frames. The common values of cross attention guidance scale λ and key-frame amplification scale τ_{key} are 2.5 and 1.2. The common values of attention sharing guidance scale $\mu_1 = 1.5$ and $\mu_2 = 2.0$. Additionally, Guided cross-attention and oriented attention-sharing, are methods delving into attention map guidance [1, 5, 9, 30] and will disrupt the latent distribution if we apply them through the entire inference stage. Thus these two methods are only applied in the first 10% and 50% of the inference steps respectively and the number of total inference steps is 50. Noted that our foreground generation should be based on the semantics of the background video, thus for each trajectory-combo, we first generate backgrounds that fit with the semantics of the bbox-prompt pairs, then continuously generate foregrounds upon the backgrounds.

Further comparisons with other bbox-based methods. Boximator [43] is a training-based approach that requires extensive training to achieve fine-grained motion control. Peekaboo [20] and TrailBlazer [24] generally add complex attention masks to multiple of spatial-, cross-, and temporal-attention, which can degrade video quality to some extent due to massive intervention in inference stage. The control module of Direct-a-Video [51] is similiar to ours, but it solely relies on the size of bbox to determine attention map edits. In contrast, our method, despite being applied within a complex layer generation pipeline, remains relatively simple in motion control but remarkably efficient and effective, as shown in Table 2.

B. Evaluation Details

B.1. Metrics Calculation

1. The reference set used for FID [37] and FVD [41] is 800 videos randomly selected from AnimalKingdom [28]. These videos will be cropped and resized to the same resolution to calculate the evaluation scores.

2. For metric CLIPSIM [33], we calculate the sample with overall conditional prompt, for example, if our layered inputs are organized as (a) "a jellyfish swimming" (b) "a carp moving" (c) "a coral reef", then the prompt for calculating the CLIPSIM will be the *overall prompt* "a jellyfish swimming, a carp moving, a coral reef".

3. For user preference regarding semantic fidelity, we invited 15 participants to select the best result from four different methods based on the following criteria: (i) Semantic Integrity – whether the generated video fully covers the requested content. For example, if the prompt specifies multiple objects but some are missing, or if a specific background is required but not generated, the result does not meet this criterion. (ii) Semantic Clarity – whether any generated objects lack clear semantic definition. If an object appears mixed with the texture of other subjects, it fails to satisfy this criterion. (iii) Overall Alignment with the Prompt – whether the generated content as a whole closely adheres to the input prompt. Based on these aspects, participants selected the result they considered the best.

4. For the metrics mIoU, AP50, Cov, and CD [20] corresponding to trajectory control capability, we separately record the bbox-object alignment between each bbox se-

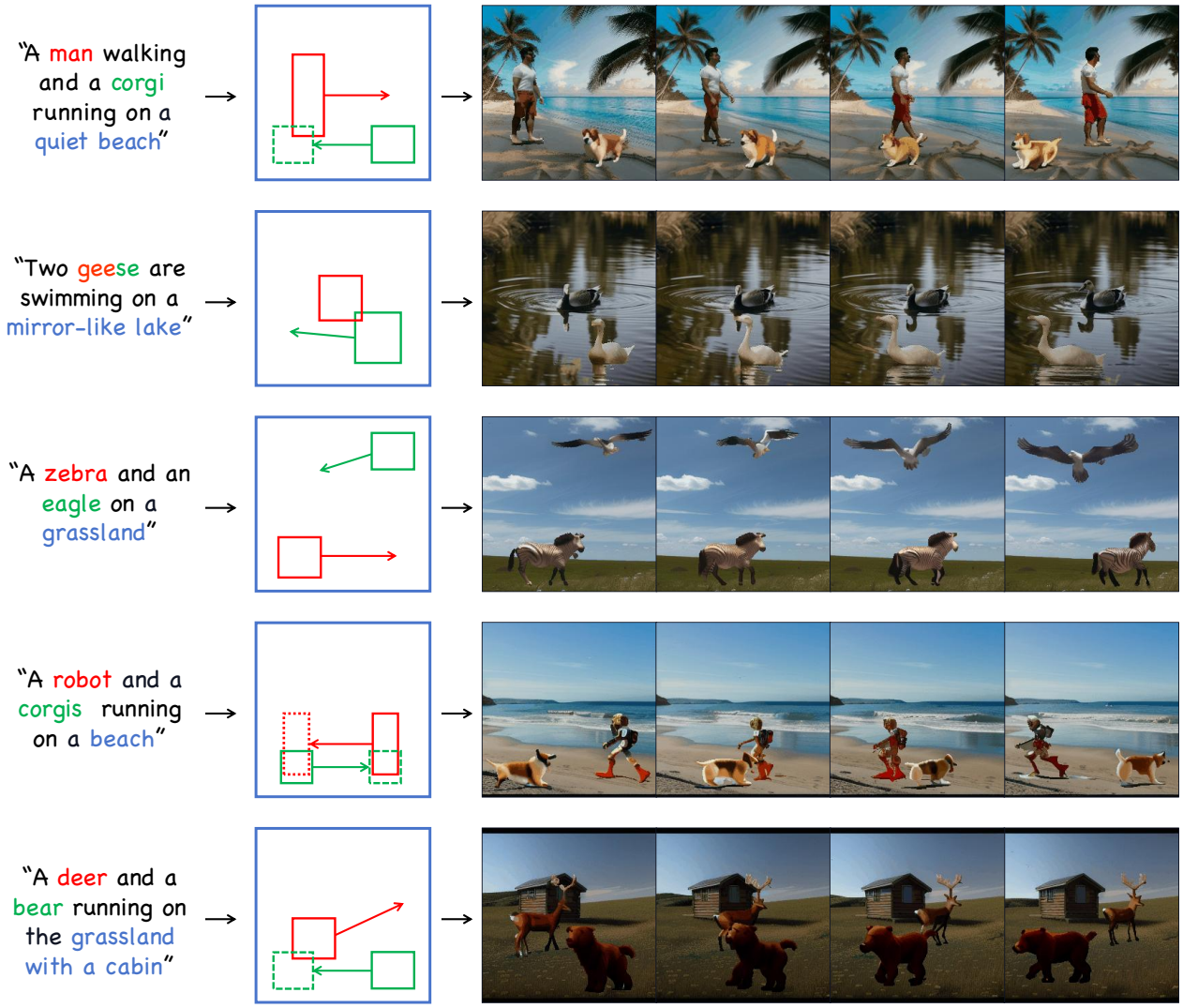


Figure 11. Additional result of multi-object motion control. The result shows five synthesized video sequences with either non-colliding or colliding multi-object motion control.

quence and its corresponding object in a trajectory-combo. Ultimately, all evaluation scores are averaged to produce the final score.

B.2. Prompts and Bboxes

Here we use the notation "Itpl($bbox_1, bbox_2, n_1, n_2$)" to represent the interpolation results for $bbox_1$ in frame n_1 and $bbox_2$ in frame n_2. And we provide some examples of the bbox-prompt combos we have used for evaluation, which are organized as {*Background Prompt, Foreground Prompt 1, Foreground Bboxes 1, Foreground Prompt 2, Foreground Bboxes 2, . . . (if any)*}, as follows:

- BG: "a coral reef in the ocean"
  FG1: "a clownfish swimming in the ocean"
  Itpl($[0.7, 0.6, 0.9, 0.8], [0.1, 0.6, 0.3, 0.8], 1, 16$)
  FG2: "a crab climbing from left to right"
  Itpl($[0.1, 0.7, 0.3, 0.9], [0.7, 0.7, 0.9, 0.9], 1, 16$)
- BG: "beautiful bright snow field with a snow mountain in the distance"
  FG1: "a polar bear walking on the snow"
  Itpl($[0.03, 0.36, 0.29, 0.66], [0.67, 0.26, 0.97, 0.58], 1, 8$)
  Intl($[0.67, 0.26, 0.97, 0.58], [0.36, 0.63, 0.67, 0.98], 8, 16$)
  FG2: "a drone flying in the sky, behind snow mountain"
  Itpl($[0.67, 0.59, 0.97, 0.88], [0.03, 0.59, 0.32, 0.88], 1, 16$)
- BG: "a cozy room with a colorful rug"
  FG1: "a corgi running in the room"
  Itpl($[0.06, 0.40, 0.31, 0.64], [0.37, 0.70, 0.64, 0.96], 1, 9$)
  Itpl($[0.37, 0.70, 0.64, 0.96], [0.66, 0.39, 0.97, 0.63], 9, 16$)
  FG2: "a ball rolling on the carpet"
  Itpl($[0.69, 0.68, 0.93, 0.94], [0.06, 0.47, 0.30, 0.71], 1, 16$)

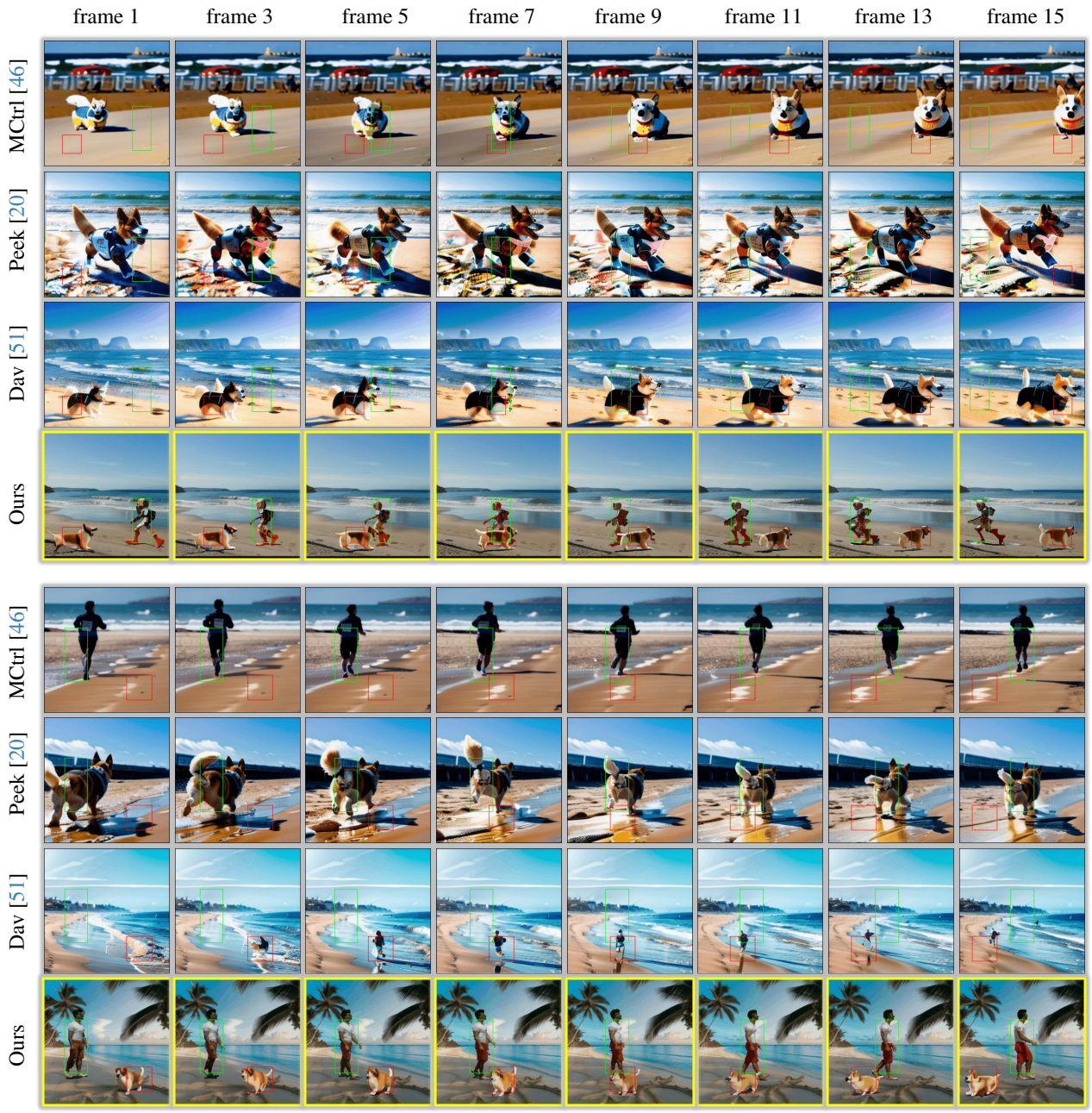


Figure 12. Further comparison on colliding object motion control with other methods. The prompt for the group above is "A robot and a corgi running on the beach", for the group below is "A man walking and a corgi running on a quiet beach".

C. More Results

C.1. Multi-Object Motion Control

In Figure 11, we present additional results of multi-object video customization, including some cases where the control information corresponds to colliding motions. In Figure 12, we present additional comparisons against other control methods. It is observed that LayerT2V demonstrates exceptional capability in generating multi-object scenes, addressing the limitations of traditional T2I/T2V models in handling multi-object generation [6]. This highlights a potential application:

Any image or video generation model trained on single-object datasets can adopt our methodology to enable support for multi-object scenes.

This paradigm eliminates the dependence on multi-object training data while enhancing the model's generative capabilities and diversity.

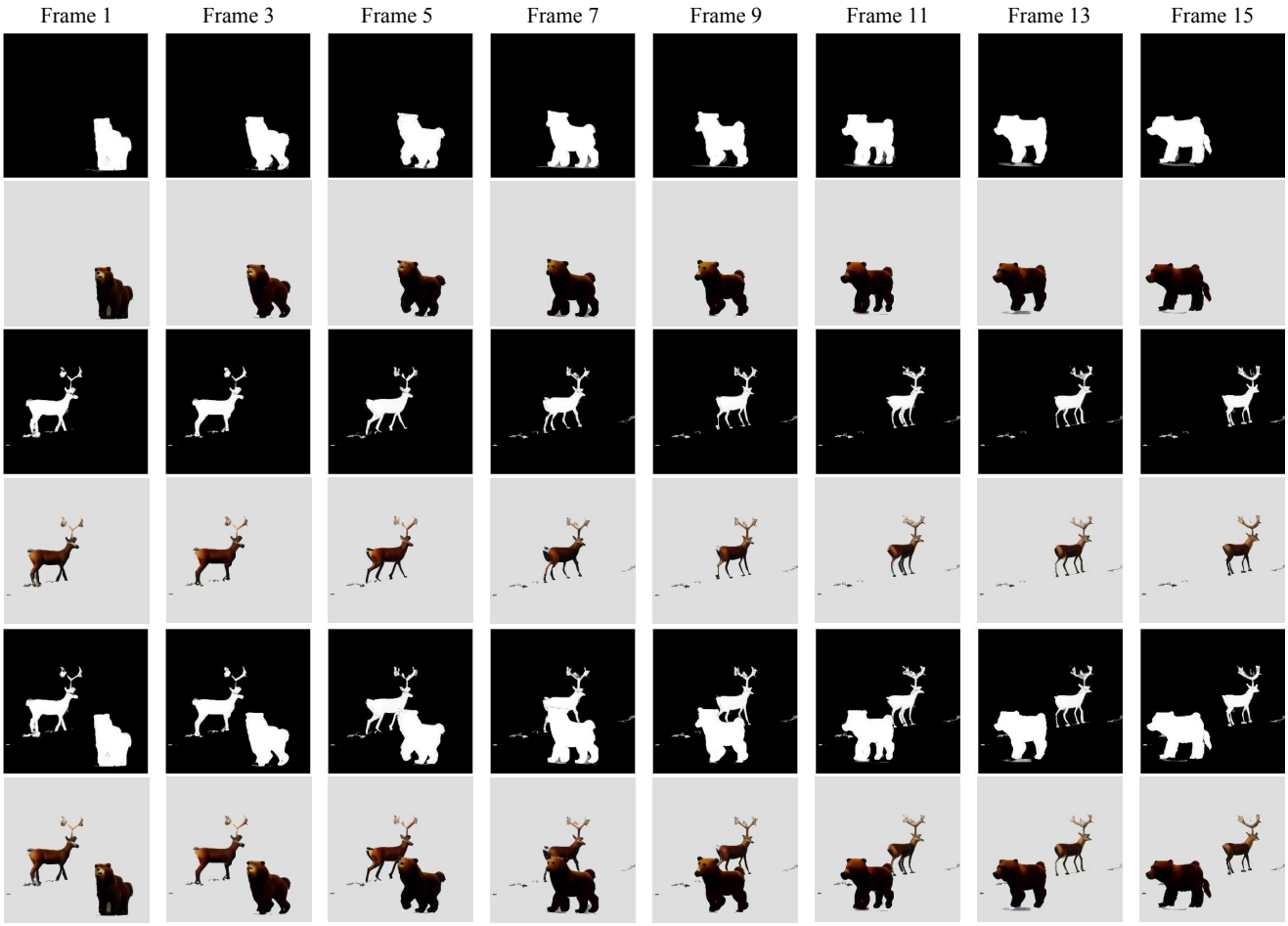


Figure 14. Visualized alpha masks of layered outputs.

Figure 15. Samples generated by base model using complicated multi-object prompts. Given prompts from top to bottm: “a bear and a corgi in a forest”, “a horse running on grassland, an eagle flying in the sky”, “a horse and a deer running on grassland”. The left part shows *semantic absence*, while the right part showcases *semantic mixing*.

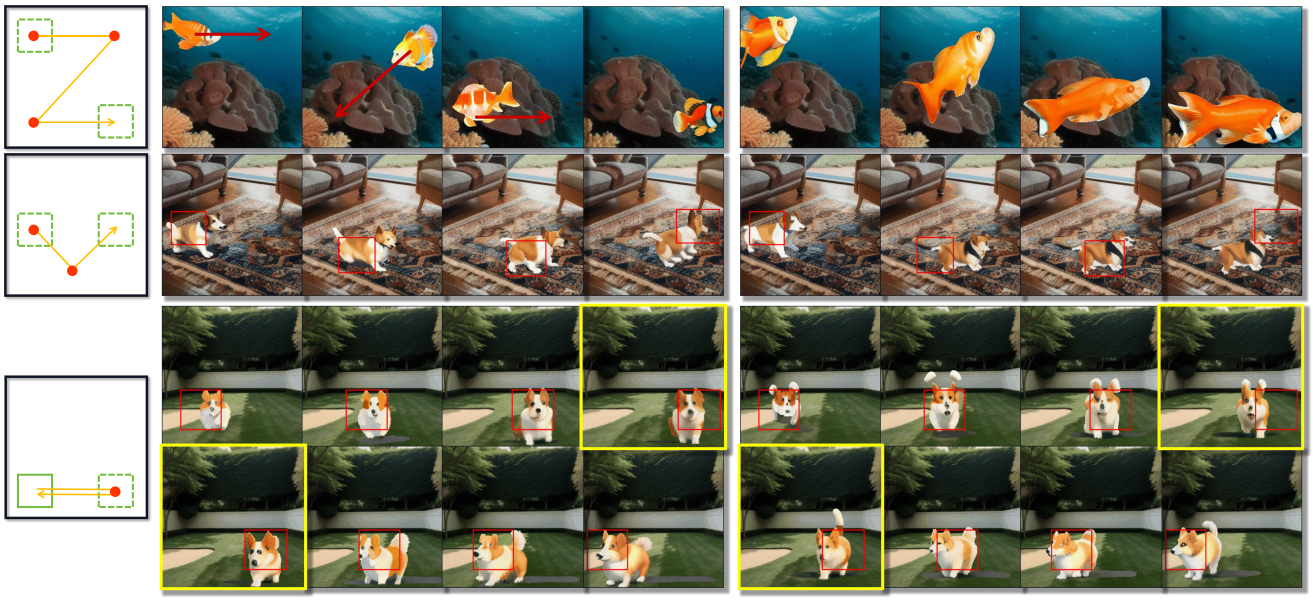


Figure 16. Additional results of Key-Frame Amplification (KFA). The prompts used in the four sets of the experiments are: "a corgi swimming in the ocean", "a corgi running in the room", "a corgi running back and forth on the lawn". Left: With KFA; Right: W/o KFA.

C.2. Alpha Masks of Foregrounds

To better illustrate the transparency relationships between multiple layers of objects, we visualize the alpha masks of each layer as well as the blended alpha mask of the foreground objects for one set of results. The visualization is presented in Figure 14.

C.3. More Results of Key-Frame Amplification

We provide additional experiments to demonstrate the importance of Key-Frame Amplification (KFA), focusing primarily on complex trajectories, which refer to polyline paths with one or more intermediate turning points. The results in Figure 16 show that our model enables more sophisticated customization of the foreground object's trajectory.

C.4. More Results of Oriented Attention-Sharing

We present additional results of the proposed Oriented Attention-Sharing (OAS) mechanism in Figure 17. OAS effectively enhances the realism of the generated foreground while ensuring better harmony with the background. Specifically, in the first example, the reflection of the generated "duck" is remarkably consistent, adding to the scene's overall realism. In the second example, our generated results appear highly realistic, whereas the toy car in the comparison appears to be floating unnaturally. In the third and fourth examples, the shadows and illumination effects in our results create a more authentic appearance, significantly elevating their aesthetic value.

C.5. More Results of Harmony-Consistency Bridge

The Harmony-Consistency Bridge (HCB) is specifically designed to handle scenarios with multiple input layers. When dealing with cases that include not only a background layer but also additional foreground layers, the subsequent layer must attend to all existing layers to maintain coherence and harmony across the composition. This ensures that interactions between layers, such as lighting, shading, and spatial alignment, are handled consistently.

Typically, the value of t_ε in Equation 13 is usually set to $0.5 \times T$, which balances the contributions from previous layers to guide the generation of the upcoming layer. This strategy ensures that each new layer integrates seamlessly with the already generated content, enhancing the realism and overall quality of the scene. Additional experimental results demonstrating the effectiveness of HCB are provided in Figure 18.

C.6. Observations of Base Model

Our base model is built upon Stable Diffusion with AnimateDiff [11]. To evaluate its capability in handling multi-object scenes, we conducted tests and present results in Figure 15. As demonstrated, the base model always exhibits the two issues discussed in Sec. 4.2. when processing complex prompts: Semantic Absence [6] (left part of Figure 15) and Semantic Mixing (right part of Figure 15). This highlights the ability of our model to surpass the limitations of T2V models in generating multi-object scenes or handling complex prompts, producing content with richer semantics.

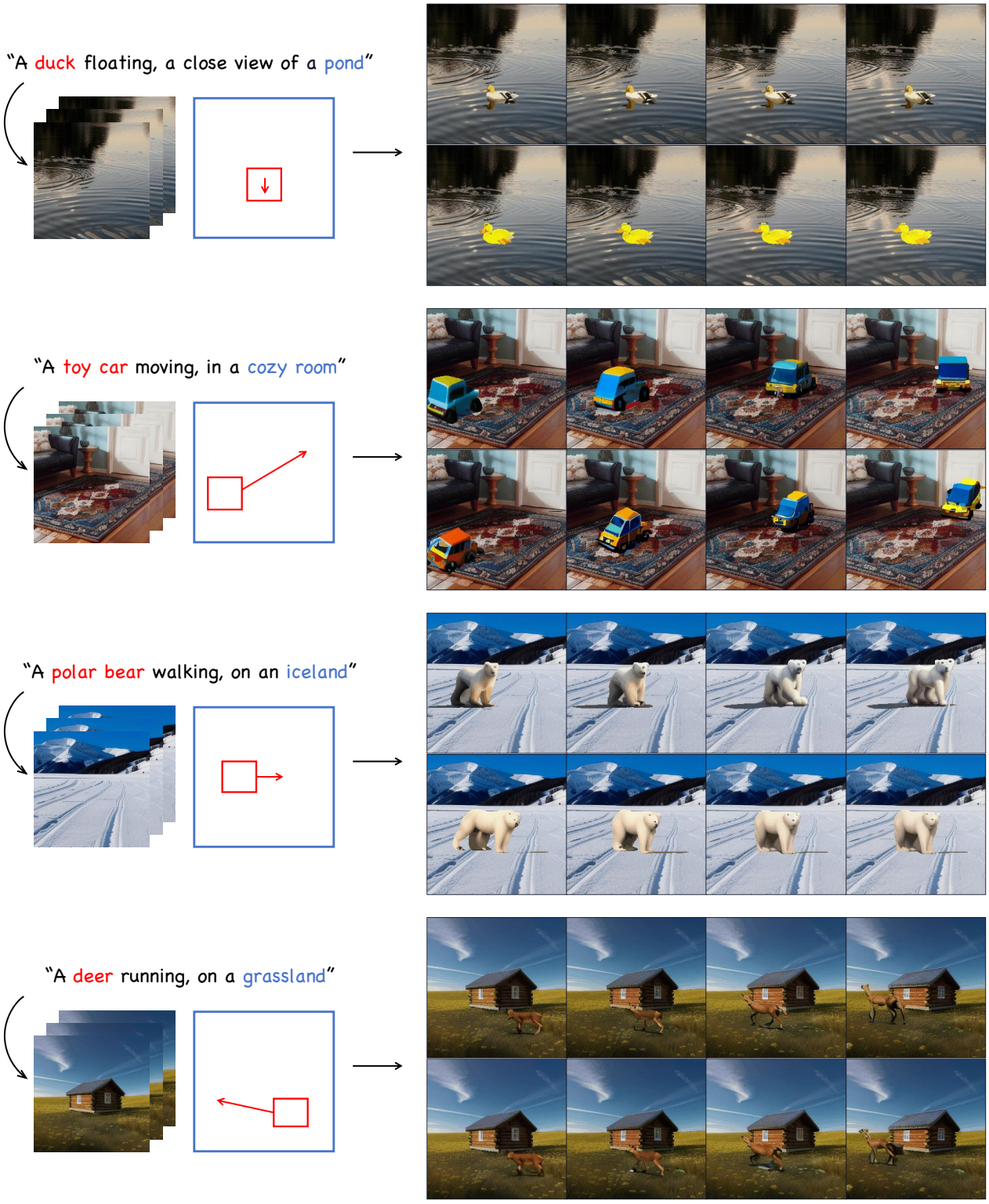


Figure 17. Additional results of the effects of Oriented Attention-Sharing (OAS). The upper row is the results of OAS and the lower one is the results for which we eliminate the OAS. We could see the shadow effects and illumination with our proposed methods is much more harmonious than those without OAS.

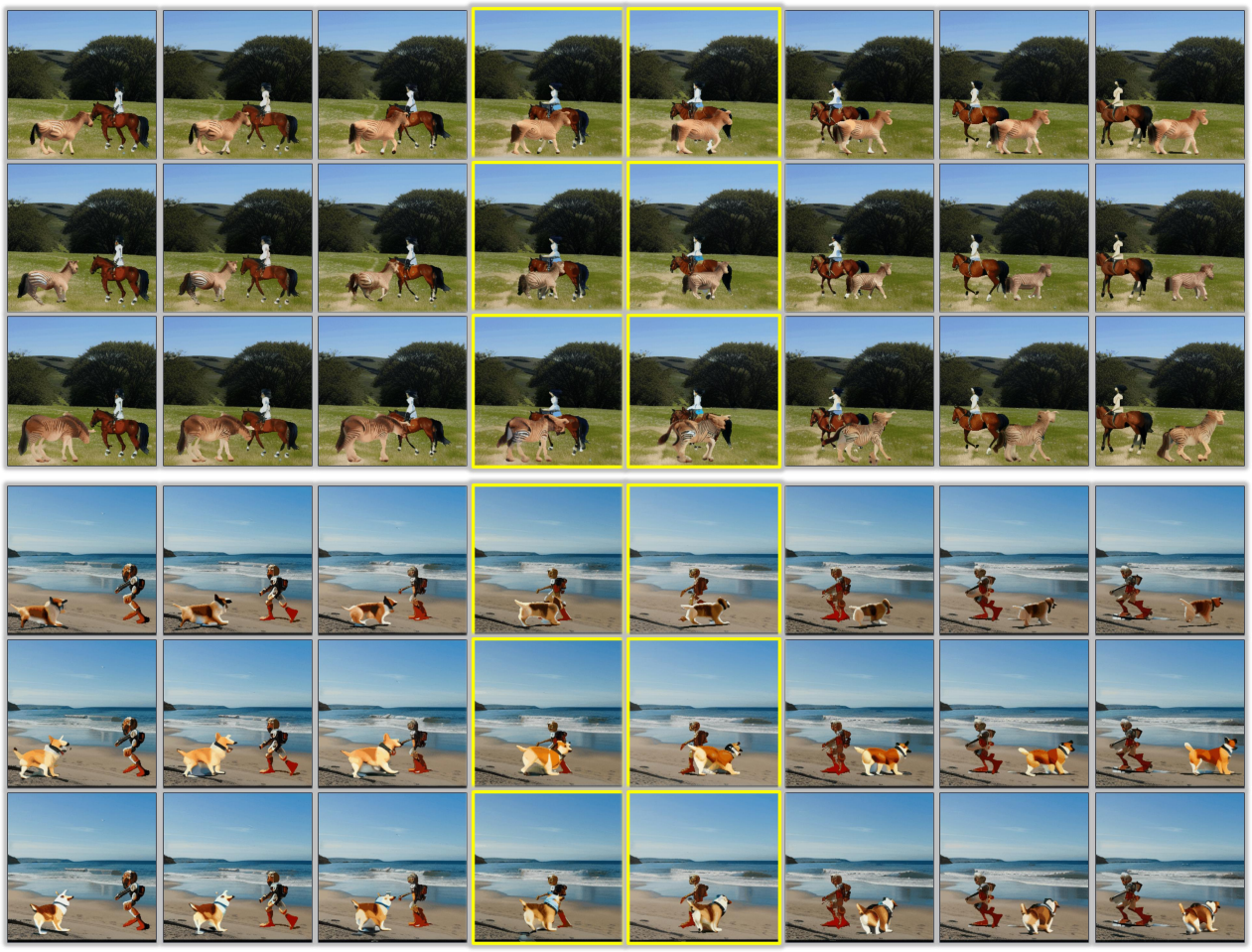


Figure 18. Additional results demonstrating the effects of the Harmony-Consistency Bridge (HCB). In each group of frames, the first row represents our results, the second row shows results generated using only the background, and the third row displays results produced using only blending. In the first group, the background is fixed with a horse, and the input is “a zebra walking on the grass”. In the second group, the background is fixed with a robot, and the input is “a corgi running on the beach”.