

# Through-The-Mask: Mask-based Motion Trajectories for Image-to-Video Generation

## Supplementary Material

### 6. Additional Results

#### 6.1. Qualitative Comparison of Masked Attention Mechanism

Fig. 6 shows qualitative comparison of generated videos for each configuration of THROUGH-THE-MASK, demonstrating the differences when applying masked cross-attention, self-attention, both, or no masked attention layers.

#### 6.2. Additional Qualitative Comparison of Motion Representation Ablation

Fig. 7 shows a qualitative comparison of the generated videos for different intermediate representation configurations of THROUGH-THE-MASK. Specifically, it compares our chosen representation, which is mask-based motion trajectories, to optical flow.

#### 6.3. Additional Qualitative Comparisons

Building upon the comparisons presented in Sec. 4.2, we provide further qualitative results comparing our approach to existing baselines. Figures 8 and 9 illustrate qualitative comparisons for DiT-based and U-Net-based models, respectively.

#### 6.4. Additional Qualitative Results

In Fig. 10, we present additional qualitative results of our method (DiT-based version) on challenging prompts, including examples featuring two or three objects, as well as a case where the main object does not appear in the input image.

#### 6.5. Effective Number of Objects

To evaluate the scalability of our method with respect to the number of objects in a video, we constructed a test set of synthetic image-prompt pairs, divided into four subsets based on the number of objects (1–4). Each subset contains 30 pairs. Running inference on this set, we observed a decline in CLIPFrame scores ( $0.971 \rightarrow 0.966 \rightarrow 0.951 \rightarrow 0.923$ ) and ViCLIP-T scores ( $0.220 \rightarrow 0.218 \rightarrow 0.216 \rightarrow 0.208$ ), with a notable drop beyond three objects.

#### 6.6. Failure Cases

Since our pipeline consists of three sequential components—prompt rewriting, motion mask generation, and video generation—errors in one stage can propagate to later stages. To evaluate the robustness of our approach under such conditions, we conducted an experiment analyzing

Setting	FVD	CF	ViCLIP-T	ViCLIP-V	AD
GT + Man.	1373.927	0.941	0.219	0.898	6.639
Gen. + Man.	1621.196	0.959	0.218	0.871	5.574
Mildly noisy + Man.	1634.298	0.948	0.218	0.867	6.289
Strongly noisy + Man.	1756.287	0.941	0.216	0.823	7.428
GT + LLM	1413.512	0.941	0.216	0.897	6.582
GT + Dropout	1509.512	0.940	0.214	0.896	6.372
GT + Wrong	1798.974	0.929	0.209	0.872	6.412

Table 5. Stage-by-stage failure case analysis.

ing potential failure cases. We randomly sampled 30 examples from SA-V-128 and evaluated three failure scenarios: (i) manually provided prompts with ground-truth (GT) mask-based motion trajectories (ideal conditions), (ii) manually provided prompts with generated noisy masks (mildly/strongly noisy), and (iii) GT masks with altered prompts (re-written, dropout-based, or incorrect). Tab. 5 summarizes the findings. We observe that strong noise in mask generation leads to greater degradation in video quality metrics, while mild noise has a minimal impact. Additionally, prompt rewriting and dropout-based modifications maintain performance close to the original GT settings, whereas incorrect prompts significantly degrade performance across all metrics.

### 7. Motion and Object-Specific Prompts Details

As described in Sec. 3.1, our pre-processing pipeline extracts a motion-specific prompt,  $c_{motion}$ , from the input text  $c$ , using a pre-trained LLM. This prompt provides a consolidated description of all motion in the scene, excluding any spatial, color, or object-specific details, and serves as a high-level guide for motion generation.

To generate the motion-specific prompt, we use Llama v3.1-8B [13] in a frozen configuration. The input prompt instructs the LLM to focus solely on motion, as shown in Fig. 11, ensuring that descriptions remain centered on movement dynamics, ignoring background information and visual characteristics of objects.

### 8. Motion-capable Objects’ Prompt Extraction Details

As described in Sec. 3.1, the pre-processing process begins with extracting motion-capable object prompts from the global prompt  $c$ . We utilize Llama v3.1-8B [13] as a frozen LLM and provide the prompt shown in Fig. 12,

which outlines the process for generating local prompts for motion-capable objects.

## 9. Inference

Given the reference image  $x^{(0)}$  and text prompt  $c$ , inference is carried out in two stages. First, the initial segmentation  $s^{(0)}$  is extracted from  $x^{(0)}$  using SAM2 [40]. Concurrently, the text prompt  $c$  is processed by a pre-trained LLM to obtain the motion-specific prompt  $c_{motion}$  and object-specific prompts  $c_{local} = \{c_{local}^{(1)}, \dots, c_{local}^{(L)}\}$  as detailed in Section 3.1. At stage 1, the image-to-motion generates motion trajectories  $\hat{s}$  conditioned on  $(s^{(0)}, x^{(0)}, c_{motion})$ . Next, in stage 2, the motion-to-video produces the final video  $\hat{x}$  by conditioning on  $(x^{(0)}, \hat{s}, c, c_{local})$  and incorporating masked attention mechanisms to ensure consistency and controllability, as described in Section 3.3. For both stages, we adopt the Classifier-Free Guidance [19] approach proposed by Brooks et al. [8]. To align precisely with their method, we treat the concatenated visual conditions as a single visual condition ( $S_I$ ) and apply the same approach to text ( $S_T$ ). We set  $S_I = 1.5$  and  $S_T = 8.5$ .

## 10. Implementation Details

As detailed above, we demonstrate the applicability of our approach to two architectures.

The first is the U-Net architecture. We follow the AnimateDiff V3 [17] design, consisting of approximately 1.4B parameters. In the second stage of motion-to-video, detailed in Sec. 3.3, we set  $K = 6$ , where  $K$  represents the number of attention blocks expanded into masked attention blocks—specifically, by adding masked self-attention and masked cross-attention into the spatial attention blocks within the U-Net’s encoder layers. The U-Net-based model was optimized using the solver suggested by [26], incorporating the DDIM diffusion solver with  $v$ -prediction and zero signal-to-noise ratio (SNR). The latter was found to be critically important to enable image-to-mask-based motion trajectory generation.

The second architecture is DiT-based. We train a DiT model following the MovieGen [38] design, containing four billion parameters. For the DiT-based model in stage two, we used  $K = 10$ , corresponding to the first 10 attention blocks out of a total of 40. The DiT-based model was optimized as described in the MovieGen paper, with Flow Matching [27], using a first-order Euler ODE solver. During inference, we adopted MovieGen’s efficient inference method by combining a linear-quadratic  $t$ -schedule, as detailed in the MovieGen paper.

For both architectures, text-to-video pre-training followed the methodology outlined in MovieGen. Across both training stages (Sec. 3.2 and Sec. 3.3), we utilized the fine-grained mask-based motion trajectories dataset described in

Sec. 3.1. The U-Net model was trained at a resolution of  $512 \times 512$ , predicting 16 frames, while the DiT model was trained at a resolution of  $256 \times 256$ , predicting 128 frames. Both models were trained on 1M video-text pairs, which were filtered as described in Sec.3.1 (with  $\tau = 0.955$ ) using 32 A100 GPUs with a global batch of 32, a constant learning rate of  $2 \times 10^{-5}$ , a warm-up period of 2000 steps, and a total of 50,000 steps. To accommodate varying input resolutions during both training and testing, each image was first resized to match the target resolution along its smaller dimension, then center-cropped to the required size.

## 11. SA-V-128 Benchmark

We introduce a balanced test set of 128 videos from the SA-V dataset [40], comprising 64 single-object and 64 multi-object cases, with an average duration of 14 seconds per video. The filtering of 128 videos, out of the full SA-V dataset, involved several steps. First, for each video, we generated a text caption using Llama v3.2-11B [13] by providing the first, middle, and last frames and asking the model to generate a caption describing the video. Next, from a closed set of categories (Animal, Architecture, Digital Art, Food, Landscape, Lifestyle, Plant, Vehicles, Visual Art, and Other), we used Llama v3.2-11B [13] to categorize each video based on these frames. We then iterated over the categories, selecting a unique category at each step and adding a related video to ensure a balanced test set. We assigned an aesthetic score and a motion score by calculating the magnitude of the optical flow extracted with RAFT [47]. After assigning captions and scores, we filtered 500 videos by iterating through each category and selecting those with the highest combined aesthetic and motion scores. From these 500 automatically filtered videos, we randomly selected 64 single-object and 64 multi-object videos. To ensure a fair comparison for shorter video settings, we also provided short captions, generated using the same methodology, extracted from frames 0 to 127 of each video. The complete benchmark is publicly available at <https://guyyariv.github.io/TTM/>.

## 12. Image-Animation-Bench

The Image-Animation-Bench comprises 2,500 videos, collected to meet high-resolution requirements and aesthetic quality thresholds. To ensure coverage of diverse visual scenarios, the dataset is divided into 16 categories: Portraits, Nature, Pets, Food, Animation, Science, Sports, City, Animation-Static, Music, Game, Animals, Industry, Painting, Vehicles, and Other.

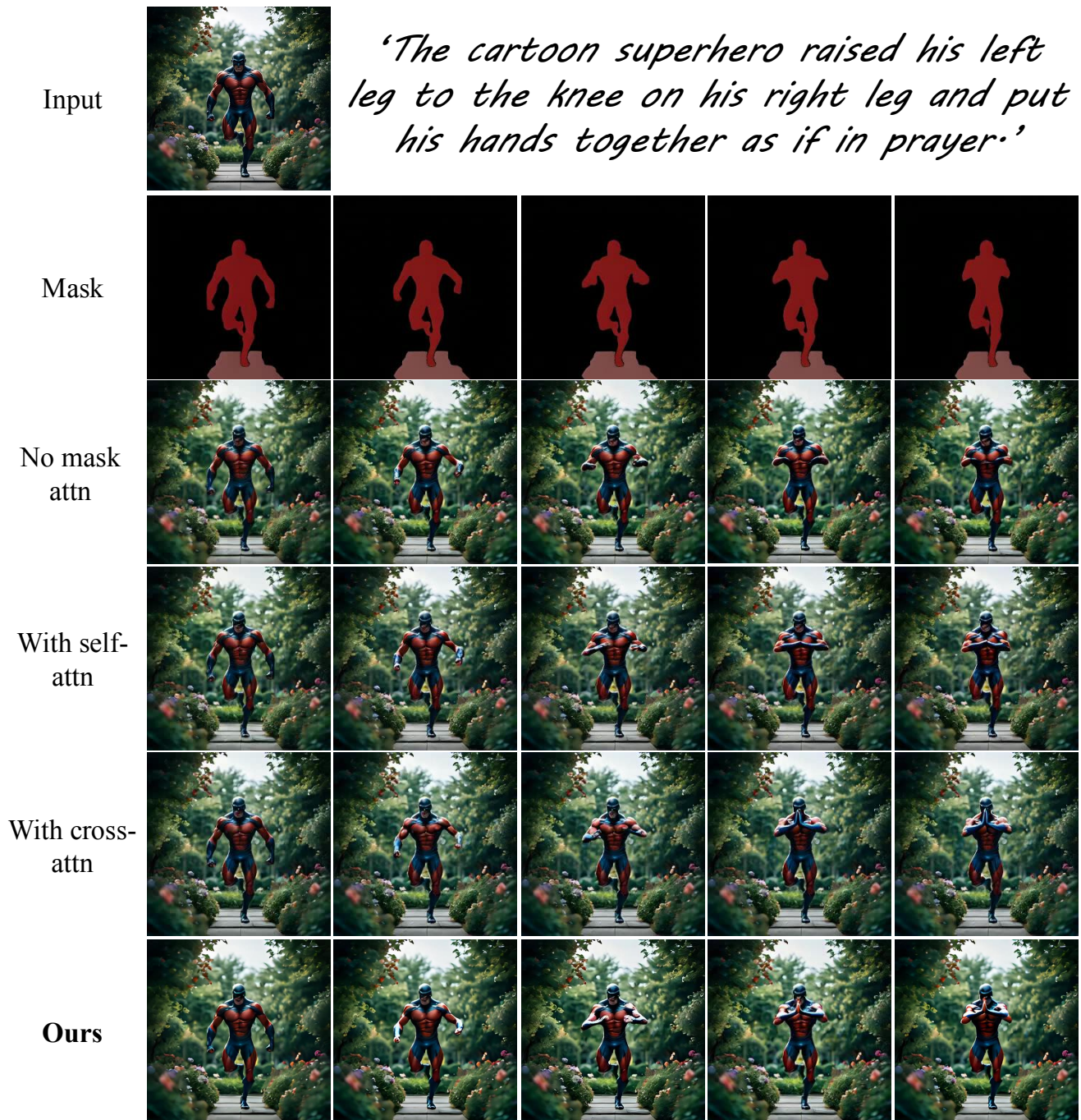


Figure 6. Qualitative comparison of generated videos for each configuration of THROUGH-THE-MASK. The results highlight differences when applying masked cross-attention (With cross-attn), self-attention (With self-attn), both (Ours), or no masked attention layers (No mask attn). Without masked attention, the cartoon superhero fails to perform a prayer. With masked self-attention, the superhero also fails, but the movement appears smoother and more consistent. With masked cross-attention, the superhero successfully performs the prayer, though his fingers turn blue. When integrating the full masked attention mechanism, the superhero performs the action correctly.

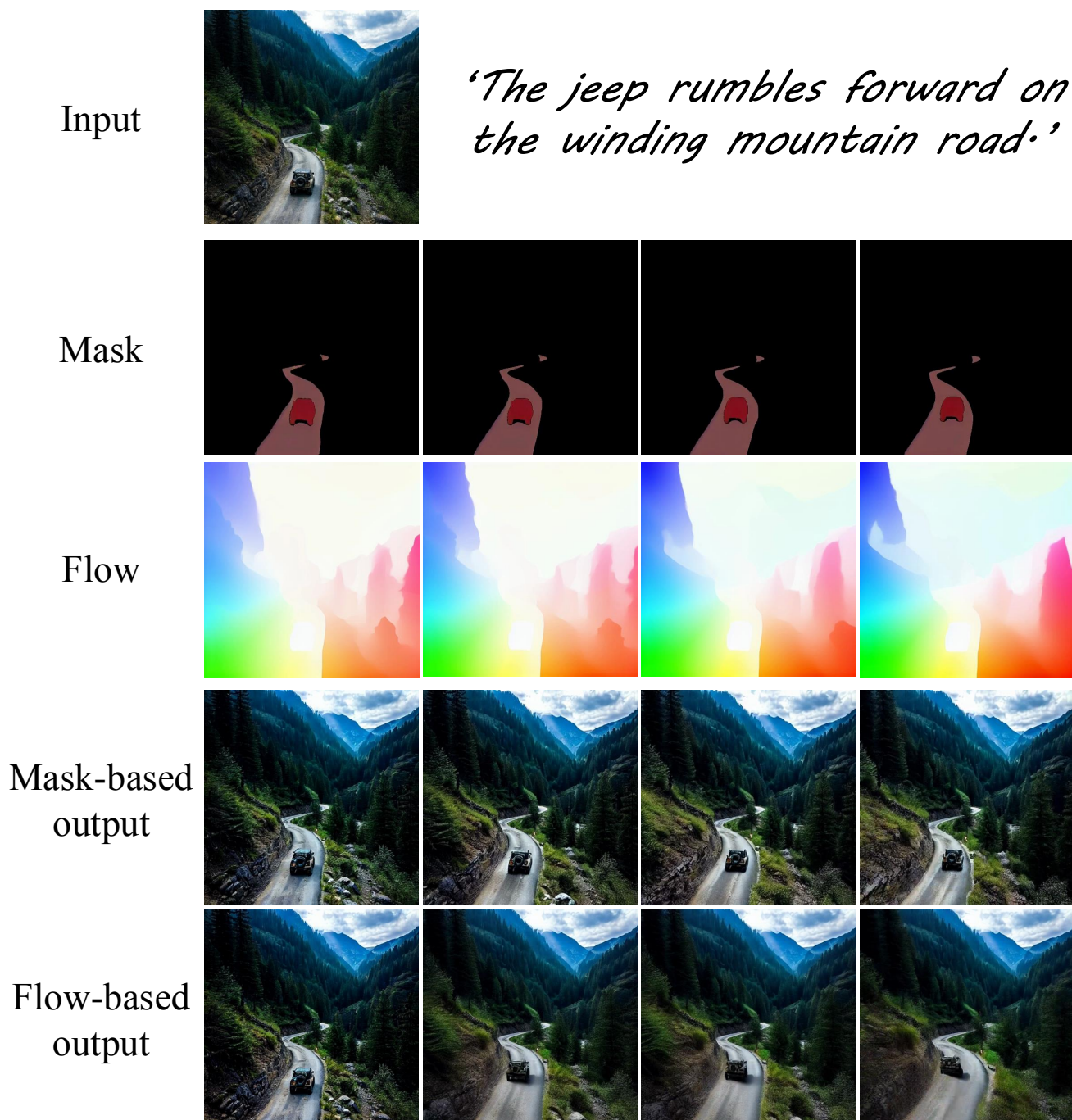


Figure 7. Qualitative comparison of generated videos using segmentation masks vs optical flow as an intermediate motion representation. The first row shows the input image and text, the second row displays the generated masks, and the third row presents the generated optical flow. The fourth and fifth rows show the generated videos, with the fourth row using our mask-based model and the fifth using our flow-based model.



Figure 8. Qualitative comparison of video generations produced by THROUGH-THE-MASK (DiT-based) and the TI2V baseline (DiT-based).

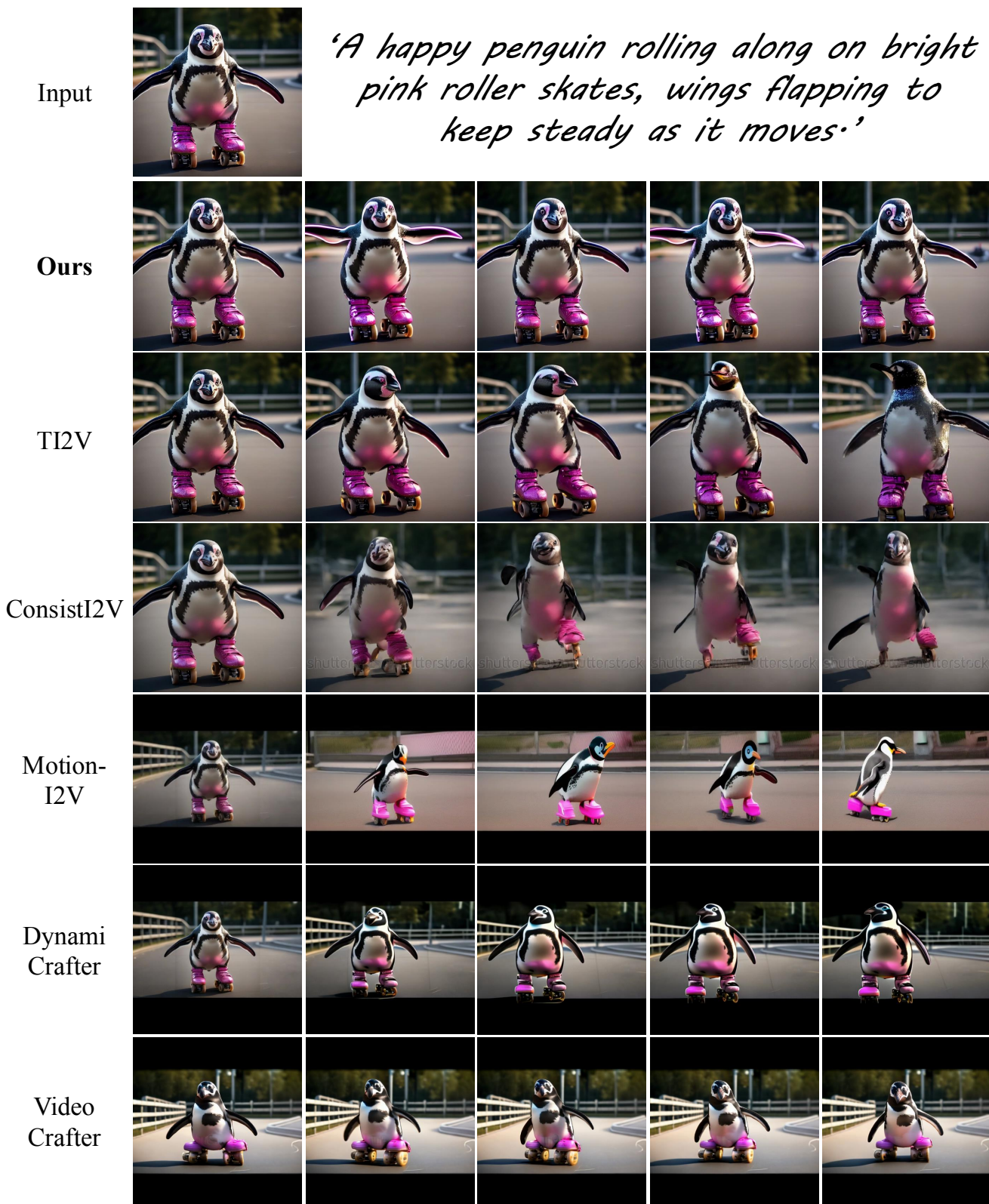
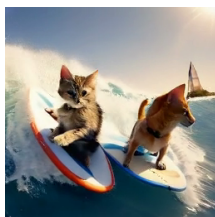
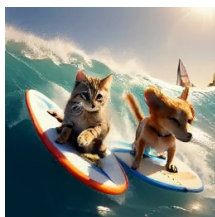
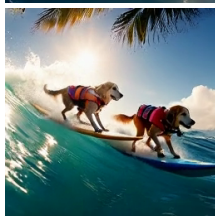
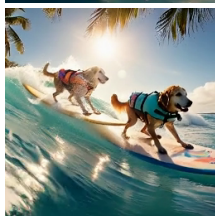
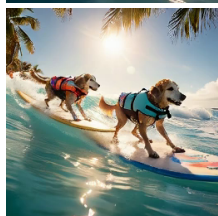


Figure 9. Qualitative comparison of video generations produced by THROUGH-THE-MASK (U-Net-based) and TI2V (U-Net-based), ConsistI2V, Motion-I2V, DynamyCrafter, and VideoCrafter.

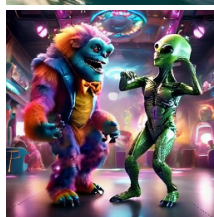
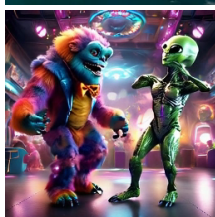
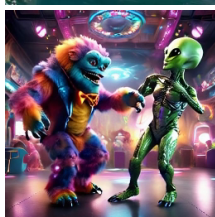
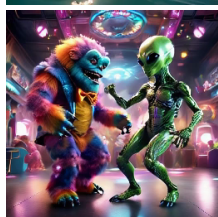
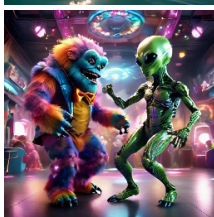
*"A cat and a dog surfing."*



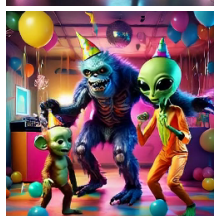
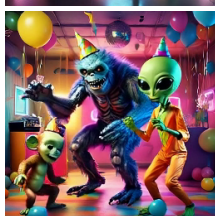
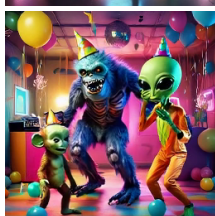
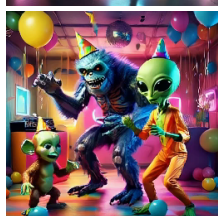
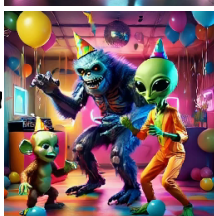
*"Two dogs surfing."*



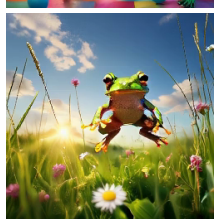
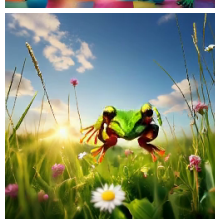
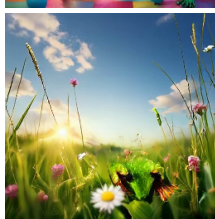
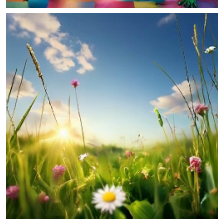
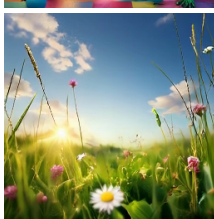
*"A monster and an alien dancing."*



*"A monster, an alien, and a monkey dancing."*



*"A frog jumping."*



(i) Input

(ii) Output

Figure 10. Additional qualitative results of our method (DiT-based version).

Task: Extract a single motion-specific prompt from the caption that describes the overall motion without including any spatial, color, size, or background details.

Format your answer like this:

Motion-specific prompt: "description of overall motion"

Examples:

Caption: "A large, red ball rolls to the right on a grassy field while a small, blue kite flies upward in the clear, blue sky."

Motion-specific prompt: "The ball rolls to the right, and the kite flies upward."

Caption: "A sleek, black car drives down a busy city street with tall buildings in the background as several pedestrians wearing bright clothing cross."

Motion-specific prompt: "The car drives down the street as pedestrians cross."

Caption: "A fluffy, white cat jumps onto a wooden table set against a plain, beige wall and knocks over a glass of water, spilling it onto the floor."

Motion-specific prompt: "The cat jumps onto the table and knocks over the glass."

Now, please provide the answer.

Caption: "{global\_prompt}"

Motion-specific prompt:

Figure 11. The input prompt used for extracting a motion-specific description from the global prompt  $c$ , designed for use with a pre-trained LLM. The prompt focuses solely on describing the overall motion, explicitly excluding details such as sizes, colors, or background elements. Here,  $c$  refers to the global prompt, which is inserted in place of {global\_prompt}.

Task: For each object mentioned in the caption, write a local prompt that describes everything about that object as mentioned in the caption.

Format your answer like this:

Answer: [[Object 1: description of object 1] [Object 2: description of object 2] ...]

Examples:

Caption: "An alien rides a horse through a field."

Answer: [[alien: A alien rides a horse through a field.]  
[horse: A horse is being ridden through a field.]]

Caption: "A dog chases a ball while a robot runs after it."

Answer: [[dog: A dog chases a ball.]  
[ball: A ball is being chased by a dog.]  
[child: A robot runs after it.]]

Caption: "An eagle flies above the mountains."

Answer: [[eagle: The eagle flies above the mountains.]]

Caption: "Two playful dogs run along the beach, with one dog on the left and the other in the middle of the frame, as waves crash onto the shore."

Answer: [[left dog: The dog runs playfully along the beach, staying closer to the dry sand.]  
[middle dog: The dog runs beside its companion, edging nearer to the waves.]]

Caption: "Three cats sit in a row on a sunny windowsill, all basking in the warm sunlight, when the cat on the right starts to move his paw."

Answer: [[left cat: The cat sits on the windowsill, soaking in the sunlight.]  
[middle cat: The cat sits on the windowsill, soaking in the sunlight.]  
[right cat: The cat sits on the windowsill, then starts to move his paw.]]

Caption: "A bustling farmers' market filled with a variety of colorful fruit stands, where a monkey is carefully picking ripe, red tomatoes while a street musician plays lively tunes on an acoustic guitar, adding a vibrant atmosphere to the scene."

Answer: [[monkey: A monkey carefully picks ripe, red tomatoes from one of the stands.]  
[musician: A street musician plays lively tunes on an acoustic guitar]]

Now, please provide the answer.

Caption: "{global\_prompt}"

Answer:

Figure 12. The input prompt used for extracting motion-capable object descriptions from the global prompt  $c$ , designed for use with a pre-trained LLM. Here,  $c$  refers to the global prompt, which is inserted in place of {global\_prompt}.