First Frame Is the Place to Go for Video Content Customization
Supplementary Material

Contents

A. Video Results

Please refer to our project page: `http://firstframego.github.io` for video results, which clearly demonstrate the effectiveness of our method and its comparison with baseline models.

B. Comparison with Two-Stage Baselines

Fig. 1 shows a comparison with a representative two-stage baseline: First composing an image layout by an image composition model like MS-Diffusion [4] and then using an I2V model for animation. Our approach offers two key advantages: 1) Our method is fully end-to-end; 2) Temporal Control via Text Prompts: More importantly, our method preserves fine-grained temporal control. For example, prompts like "a shark joins the party later" are faithfully realized. In contrast, two-stage pipelines lose this capability, as the fixed first frame dictates the entire spatial layout, limiting temporal flexibility.

Figure 1. Comparison with MS-Diffusion + I2V.

C. Details about Training and Testing Set

C.1. Training Dataset Curation Details

Our training corpus is sourced from three datasets: one randomly selected folder from HOIGen-1M [1] ($\approx$ 2,000 clips), all five Veo 3 demonstration videos [5], and 200 licensed short videos. This yields 2,205 candidate clips.

We manually curate the data to select videos with clearly separable foregrounds, humans or manipulable objects, set against uncluttered backgrounds. Only clips depicting cleanly segmentable single- or multi-object interactions are retained.

This filtering results in 50 high-quality training examples (Figure 4), distributed across four scene types: human–object interaction (60%), human–human interaction (14%), element insertion (20%), and robot manipulation (6%).

Training Data Processing. After curation, all clips are standardized to 81 frames for consistent training. From each video, we extract the first frame as a reference image and manually tag all foreground entities of interest, e.g., *cake*, *party hat*, *male presenter*, *mouse*. Using a prompt-to-prompt workflow with Gemini-2.5-Pro, we then perform:

- **Object Extraction:** Apply Prompt 2 to generate high-fidelity renditions of each tagged entity, preserving their original appearance and scale. We refine results using SAM 2 or Adobe Photoshop to isolate each object as an RGBA layer.
- **Background Cleanup:** Use Prompt 3 to produce a clean companion image with all tagged objects removed, yielding a pristine background plate.

This paired set of object cut-outs and object-free backgrounds forms the compositional basis of the training first frame.

Caption Generation. We use Gemini-2.5-Pro to generate rich, element-aware captions for each training sample, based on the individual object cut-outs, clean background plate, and the full 81-frame video. These inputs are paired with a structured prompt template (Fig. 5) to ensure consistency and relevance.

Element Composition for First Frame. For each training clip, we synthesize a 1280×720 reference canvas: all foreground cut-outs are vertically tiled on the left half, while the clean background is centered on the right (see Fig. 4). This composite serves as both the conditioning input and the initial frame, guiding the video generation model to blend the elements into a coherent sequence.

C.2. Test Set Curation

We manually curated a diverse test set of foreground objects and backgrounds from our self-collected images. Each object was segmented using SAM 2 or Adobe Photoshop and saved as an RGBA cut-out. These cut-outs were then composited with their respective backgrounds on a 1280×720 canvas, following the same layout used in training.

For each object-background pair, we drafted an initial prompt and refined it using Gemini-2.5-Pro with the template shown in Fig. 6. This process produced 50 high-quality prompts paired with composite reference images, forming our final test set.

Object Extraction Task Prompt Template

Prompt – Given the input image, extract the subset `{IDENTIFIED OBJECT}` (i.e., only the specified foreground objects)— return them *alone* with **no resizing, compression, or background** so the output resolution exactly matches the original image.

Figure 2. Prompt for extracting identified foreground objects using a unified VLM.

Object Removal Task Prompt Template

Prompt – Given the input image, **remove** the subset `{IDENTIFIED OBJECTS}` entirely. Return the edited image *only*—it must preserve the source resolution (no scaling or compression) and contain neither the specified objects nor any artifacts of their removal.

Figure 3. Prompt and specifications for the object removal task.

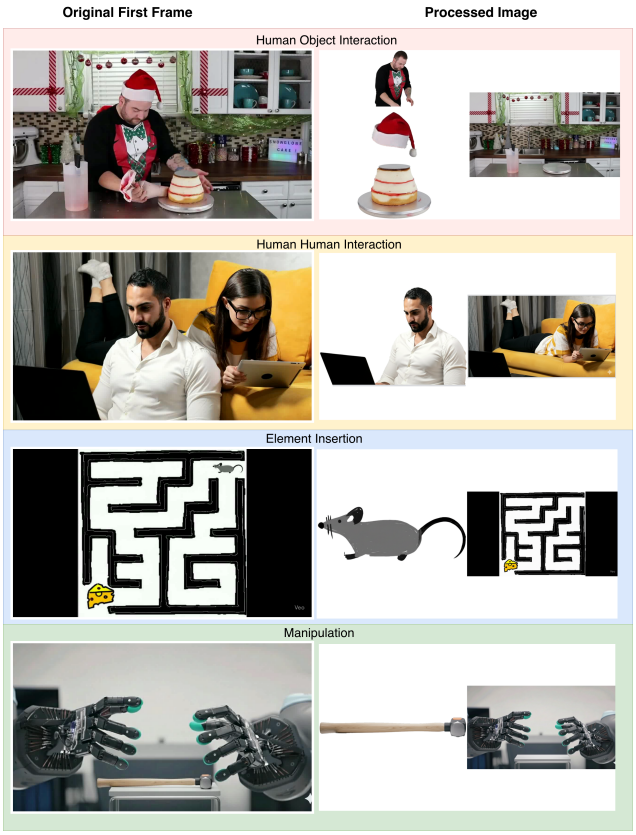


Figure 4. Our training dataset comprises four categories: human–object interaction (60%), human–human interaction (14%), element insertion (20%), and robot manipulation (6%).

D. Details about User Study

To ensure a smooth user study and annotation experience, we developed an HTML-based interface for participants to annotate and submit data. In this section, we describe the hiring platform, the job posting, and the design of the annotation interface.

D.1. User Study Platform

We recruit participants through Prolific,[1] a research platform designed for user studies. Prolific offers an AI user study beta program that targets participants with experience in generative AI annotation.

To ensure quality, we apply screening filters to select participants with prior video annotation experience and fluent English proficiency, as understanding nuanced textual prompts is crucial for this task.

We hire 40 participants, each tasked with annotating five video sets, where each set contains generated outputs from four different models. The annotation process takes approximately 15 minutes per participant. Each is compensated $5.50, reflecting the expected time and effort.

[1] https://www.prolific.com/

Training Data Prompt Generation Prompt Template

Task Description

You are given a video and several images. Generate a *descriptive caption* for the video that prominently features the components shown in the images. Wrap your final text in `<caption>...</caption>` tags. The caption must highlight the significance and role of these components throughout the video, while omitting filler such as "The scene unfolds with a whimsical and heartwarming narrative, emphasizing the simple joys of life through the Teddy Bear's endearing actions".

Examples of Descriptive Captions

1. Film quality, professional quality, rich details. The video begins to show the surface of a pond, and the camera slowly zooms in to a close-up. The water surface begins to bubble, and then a blonde woman is seen coming out of the lotus pond soaked all over, showing the subtle changes in her facial expression.
2. A professional male diver performs an elegant diving maneuver from a high platform. Full-body side view captures him wearing bright red swim trunks in an upside-down posture with arms fully extended and legs straight and pressed together. The camera pans downward as he dives into the water below.

Figure 5. Prompt template used to generate captions for our training data.

Video-Prompt Enhancement Output

Task Description

You will be given a prompt and several images for video generation. You task is to make the prompt richer in description so the model can understand better. Enclose your caption within `<caption></caption>` tags. The caption must emphasize the significance and role of these components (and some description of each component) throughout the video. Your caption should exclude unnecessary information such as "The scene unfolds with a whimsical and heartwarming narrative, emphasizing the simple joys of life through the Teddy Bear's endearing actions".

Example of a Descriptive Caption

1. Film quality, professional quality, rich details. The video begins to show the surface of a pond, and the camera slowly zooms in to a close-up. The water surface begins to bubble, and then a blonde woman is seen coming out of the lotus pond soaked all over, showing the subtle changes in her facial expression.
2. A professional male diver performs an elegant diving maneuver from a high platform. Full-body side view captures him wearing bright red swim trunks in an upside-down posture with arms fully extended and legs straight and pressed together. The camera pans downward as he dives into the water below.

Prompt to Optimize

{Insert your test prompt to optimize here}

Figure 6. Prompt template for test prompt enhancement.

Our recruitment post and task instructions are shown in Figure 7.

D.2. User Interface Details

Participants first arrive at a login screen, where they enter their unique Prolific ID to match their responses with task-completion records. After authentication, they are presented with the textual prompt used to generate the videos, along with a composite reference image showing the required foreground objects and background. Below, four candidate videos are displayed in a randomized order to eliminate position bias. Participants then rank the videos based on overall quality, as shown in Figure 8a.

Next, participants scroll down to rate each video on three criteria, Object Identity, Scene Identity, and Overall Quality, using a 5-point Likert scale (Figure 8b).

E. More Training and Inference Details

We train LoRA modules of rank 128 for both high- and low-noise regime transformers in the base model Wan2.2-I2V-A14B. Training videos are resized to a resolution of 1344×768 with 81 frames. We use a batch size of 4 and optimize

Annotation Task Instructions

You will be presented with five sets of short, AI-generated videos (*5 s, no audio*).

Each set contains:

- **Prompt** – textual description of the intended video (scene, objects, motion).
- **Reference Image** – split into two halves:
  - *Left side*: foreground objects that should appear in the video.
  - *Right side*: background scene to be integrated with the objects.
- **Generated Videos (4 total)** – four model outputs attempting to fuse the objects with the background.

Your Task for Each Set

Step 1: Overall Ranking

- Watch *all four* videos carefully.
- Rank them from best to worst based on overall quality and faithfulness to the prompt.
- Assign unique ranks (1 = best, 4 = worst).

Step 2: Aspect Ratings

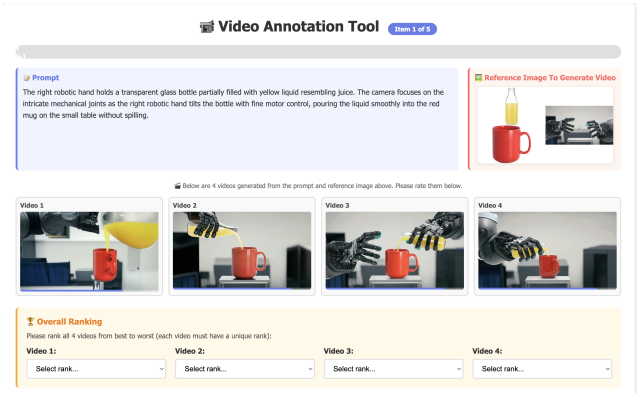
After ranking, rate each video on a 1–5 scale (1 = very poor, 5 = excellent):

- **Object Identity** – How well do objects retain their identity?
- **Scene / Background Identity** – How well is the background preserved?
- **Video Quality** – Overall realism and temporal coherence.

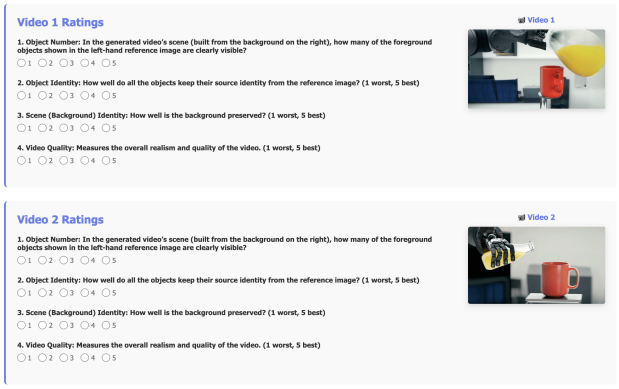
Notes

- Evaluate **all four videos** in every set *before* submitting answers.
- There are five sets in total (20 videos).

Figure 7. Recruitment post for our user study.



(a) Users rank the overall quality of the four candidate videos.



(b) Users rate three specific aspects with a Likert scale 1-5.

Figure 8. Web-based annotation interface used in our user study. Part (a) collects a global quality ranking, while part (b) gathers detailed aspect-wise ratings for each video.

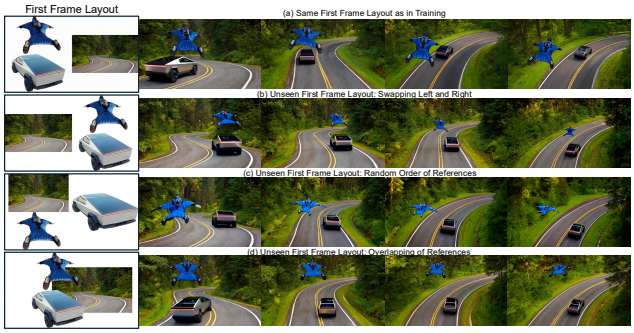


Figure 9. Generalization to spatial layouts in the first frame.

Figure 10. References from different sources.

Model	CLIP-I↑	DINO-I↑	CLIP-cap↑	Motion Smoothness ↑	Dynamic Degree↑	Aesthetic Quality↑	Imaging Quality↑
Wan2.2-I2V-A14B	0.66	0.42	33.2	0.96	9.75	0.82	0.61
VACE	**0.68**	**0.46**	33.6	0.97	7.78	**0.92**	0.65
SkyReels-A2	0.66	0.43	33.1	0.96	12.93	0.91	0.72
Ours	0.67	**0.46**	**34.00**	**0.98**	**14.64**	0.85	**0.73**

Table 1. Concept and VBench Automatic Quantitative Metrics

with AdamW [2], setting the learning rate to 1×10^{-4}, $\epsilon = 1 \times 10^{-10}$, and a weight decay of 3×10^{-2}.

During inference, videos are generated at a resolution of 1280×720 with 81 frames, following the standard output format of Wan2.2-I2V-A14B based models.

F. Generalization to First-Frame Layouts

Although training uses a fixed first-frame layout (a), our model can generalize to unseen layouts in some cases. As shown in Fig. 9, we evaluate three novel layouts (b), (c), and (d), beyond the training layout of cut-outs on the left and background on the right. The results suggest that our model interprets the first frame contextually rather than relying solely on the seen training layout.

G. Visual Consistency Across Generated Videos from Different Reference Sources

For all test results presented in the paper, the reference inputs are drawn from different sources rather than a single video. Visual consistency is maintained in the generated videos due to the pre-trained models' learned priors. For instance, in Fig. 10, fine-grained shadows cast by a hand and bottle are correctly rendered on the teddy bear (shown by arrows).

H. Automatic Quantitative Metrics

In Table 1, we show standard automatic concept and VBench quantitative metrics: CLIP-I, DINO-I, CLIP-cap (text alignment), Motion Smoothness, Dynamic Degree (motion intensity), Aesthetic Quality, and Imaging Quality, to compare with baselines. These standard metrics further validate the effectiveness of our method.

I. Explanation of the Transition Phrase

The transition phrase <transition> (e.g., "ad23r2 the camera view suddenly changes") serves as a unique trigger in the text prompt. Paired with LoRA training, it enables the base model to learn to invoke latent abilities for scene cuts and reference fusion when encountered during inference. The choice of trigger can be arbitrary, as long as it is unique. This design is inspired by the use of unique trigger phrases in DreamBooth [3], but serves a fundamentally different purpose.

References

[1] Kun Liu, Qi Liu, Xinchen Liu, Jie Li, Yongdong Zhang, Jiebo Luo, Xiaodong He, and Wu Liu. Hoigen-1m: A large-scale dataset for human-object interaction video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24001–24010, 2025. 1

[2] Ilya Loshchilov, Frank Hutter, et al. Fixing weight decay regularization in adam. *arXiv preprint arXiv:1711.05101*, 5(5): 5, 2017. 5

[3] 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. 5

[4] Xierui Wang, Siming Fu, Qihan Huang, Wanggui He, and Hao Jiang. Ms-diffusion: Multi-subject zero-shot image personalization with layout guidance. *arXiv preprint arXiv:2406.07209*, 2024. 1

[5] Thaddäus Wiedemer, Yuxuan Li, Paul Vicol, Shixiang Shane Gu, Nick Matarese, Kevin Swersky, Been Kim, Priyank Jaini, and Robert Geirhos. Video models are zero-shot learners and reasoners. *arXiv preprint arXiv:2509.20328*, 2025. 1