Supplementary Material

Overview. This supplementary material provides additional technical details, extended analyses, and qualitative results for *VideoCanvas*. We also include broader comparisons and ablations that complement the main paper. More results are available on our **Project Page**: https://onevfall.github.io/project_page/videocanvas/. The content is organized as follows:

- **Section A** (Related Work) provides a comprehensive review of related literature.
- **Section B** (Introduction of the Base Text-to-Video Generation Model) reviews the base text-to-video diffusion backbone and key design choices.
- **Section C** (Implementation Details) describes training strategy, compared methods, and evaluation protocols for reproducibility.
- **Section D** (VideoCanvasBench Construction Details) details dataset curation, task definitions, and annotation/licensing notes.
- **Section E** (More Analysis and Results) provides deeper analyses (e.g., padding robustness) and additional quantitative/qualitative results.
- **Section F** (Applications and Qualitative Results) showcases diverse applications and side-by-side visual comparisons across tasks.

A. Related Work

A.1. Arbitrary Spatio-Temporal Video Completion

Controllable video generation aims to synthesize content that adheres to user inputs beyond a simple text prompt. Existing approaches are often constrained by rigid, task-specific formats, such as conditioning only on a first frame [14, 15, 27, 40, 48], on a short initial sequence [1, 54], or on structural inpainting and outpainting [2, 19, 50, 55, 61]. Conceptually, these represent special cases of the broader challenge of video completion, yet prior work has treated them as separate sub-tasks, each requiring specialized solutions. Several recent works [7, 9, 47] explore aspects of multi-frame conditioning or temporally flexible video generation. SEINE [7] focuses on short-to-long video transitions and temporal prediction, using a temporal mask modeling objective on continuous frame sequences. Although it enables multi-frame temporal reasoning, SEINE does not support spatially irregular conditioning. VDT [35] takes a complementary approach through masked video modeling on dense spatio-temporal token grids. It relies on a spatial-only VAE (i.e., no temporal compression), producing per-frame token maps over which bi-directional prediction is performed. MCVD [47] employs masked conditional video diffusion for prediction, generation, and interpolation tasks, but operates on pixel space without leveraging modern causal VAEs. LDMVFI [9] applies latent diffusion models specifically to video frame interpolation between given keyframes, but is constrained to dense temporal sequences and does not address arbitrary spatial-temporal positioning. In contrast, we focus on the task of *arbitrary spatio-temporal video completion* under modern latent video diffusion models with causal video VAEs, addressing the pixel-frame ambiguity challenge that these prior works do not encounter.

A.2. Paradigms for Video Conditioning

Achieving arbitrary spatio-temporal control requires a robust conditioning mechanism. Existing approaches can be broadly categorized into three paradigms, each with distinct limitations when applied to our task. *Latent Replacement* [17, 27] directly overwrites latent slots with conditional content, but suffers from train-inference mismatch when applied beyond first-frame conditioning, often causing motion collapse. *Channel Concatenation* [49, 56] and adapter-based methods [22, 36, 58] fuse conditions via concatenation or lightweight encoders, yet require costly VAE and DiT retraining to handle the zero-padding needed for pixel-frame-aware control. *In-Context Conditioning (ICC)* [13, 16, 18, 20, 33, 53], pioneered by OminiControl [42] for images and extended to video by FullDiT [24] and UNIC [57], offers a parameter-free alternative by treating conditions as tokens in a unified sequence. While promising, prior ICC methods struggle with the *pixel-frame ambiguity* introduced by causal VAEs, limiting precise temporal alignment.

Building on ICC, we are the first to enable pixel-frame-level arbitrary spatio-temporal video completion under frozen causal VAEs. Our key innovation is *Temporal RoPE Interpolation*, which assigns fractional temporal positions to conditional tokens, achieving sub-latent precision without VAE retraining. Combined with a hybrid conditioning strategy, our approach unlocks ICC's full potential for fine-grained spatio-temporal control.

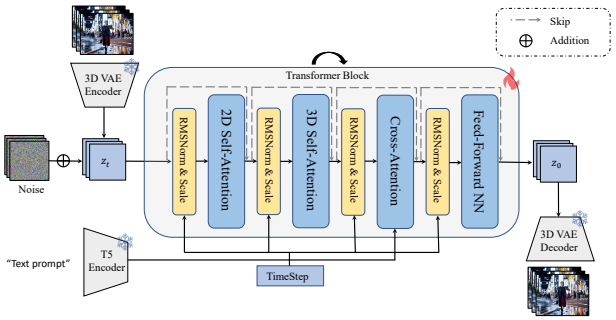


Figure S9. **Overview of the base text-to-video generation model.**

B. Introduction of the Base Text-to-Video Generation Model

We use a transformer-based latent diffusion model [38] as the base T2V generation model, as illustrated in Fig. S9. We employ a 3D-VAE to transform videos from the pixel space to a latent space, upon which we construct a transformer-based video diffusion model. Unlike previous models that rely on UNets or transformers, which typically incorporate an additional 1D temporal attention module for video generation, such spatially-temporally separated designs do not yield optimal results. We replace the 1D temporal attention with 3D self-attention, enabling the model to effectively perceive and process spatiotemporal tokens, thereby achieving a high-quality and coherent video generation model. Specifically, before each attention or feed-forward network (FFN) module, we map the timestep to a scale, thereby applying RMSNorm to the spatiotemporal tokens.

C. Implementation Details

C.1. Training strategy

The model is fine-tuned for 20k steps on a curated high-quality video dataset comprising diverse scenes and motion patterns (384×672 resolution, 5 seconds per clip), using the Adam optimizer with a learning rate of 5×10^{-5} and a batch size of 32 on 32 GPUs.

At each iteration, 20 frames are randomly sampled from a source video to serve as temporal anchors. From each anchor frame, we extract a spatial region by cropping a patch covering between 20%–100% of the original frame size. This unified training strategy ensures that the model encounters a diverse spectrum of conditioning scenarios, ranging from sparse local patches to nearly complete frames, and from early anchors to late anchors. Such exposure allows the model to learn arbitrary spatio-temporal conditioning in a single framework.

C.2. Details of Compared Methods

This section provides additional details on the methods compared against our approach, including (i) existing state-of-the-art models across different tasks, (ii) the conditioning paradigms used for fair evaluation, and (iii) the alignment strategies included in our ablation studies.

C.2.1. Compared SOTA Methods

We evaluate our method against state-of-the-art video generation systems spanning multiple task settings:

- **Image-to-Video (I2V).** We compare with CogVideoX-1.5 [56] and HunyuanVideo [27], which represent strong image-to-video generators.
- **First–Last-Frame-to-Video (FLF2V).** For FLF2V generation, we include CogVideoX-FT [12] and Sci-Fi [6], both of which are designed to synthesize temporally coherent sequences conditioned on the first and last frames.
- **First–Middle–Last-Frame-to-Video (TF2V).** To evaluate tri-frame conditioning, we decompose the task into two FLF2V subproblems—first→middle and middle→last—and stitch the two generated segments into a complete sequence. Both subproblems are solved using CogVideoX-FT [12] and Sci-Fi [6].
- **First–Last-Patch-to-Video (FLP2V).** Since this task requires completing missing regions from spatial patches, we first convert the patch inputs into full-frame images using FLUX [28], and then employ Wan-FLF2V-14B [49] to generate the full video sequence conditioned on the completed first and last frames.
- **Video Inpainting.** For temporal inpainting, we compare with VACE [22] and ProPainter [61] for reconstructing missing or corrupted intervals.
- **Video Outpainting.** For spatial and temporal outpainting, we benchmark against VACE [22] and M3DDM [11], which are designed to extrapolate beyond the original field of view or temporal extent.

C.2.2. Compared Conditioning paradigms Setting

For fair comparison of different conditioning paradigms, we re-implement other two representative paradigms on the same base model (Fig. 2b), following the references used in the main text:

- **Latent Replacement** [17, 27]. For a given conditional frame, the corresponding latent tokens are overwritten with VAE-encoded ground-truth latents. Training applies a masked loss only to non-conditional regions, while conditional regions are assigned timestep 0.
- **Channel Concatenation** [49, 56]. Condition frames are encoded into latents, assembled into a zero-padded latent sequence, and concatenated with the noisy latent sequence along the channel dimension. A learnable projection layer then restores the embedding dimension. In our implementation, concatenation is applied *after patchification*, as this setting empirically yields the best results; applying it before patchification leads to degraded visual quality. The tradeoff is that after-patchify concatenation substantially increases the channel dimensionality, resulting in a projection layer with $\sim$16.6M trainable parameters. Thus, while this design enriches the conditioning signal and improves learning, it comes at the cost of significantly more parameters compared to the other paradigms.
- **In-Context Conditioning (Ours)** [24, 42]. Our method encodes condition frames into clean latent tokens and concatenates them with the noisy sequence along the token dimension. Temporal alignment is achieved with our RoPE Interpolation strategy (Sec. 2.5). The loss is applied only to noisy tokens, while conditional tokens are assigned timestep 0. This design requires no additional trainable

parameters.

All paradigms are trained under identical settings and restricted to the same set of conditionable frames defined by the VAE stride, ensuring a rigorous and controlled comparison.

C.2.3. Ablation Alignment Strategies

Modern causal video VAEs do not encode frames independently; instead, every latent token represents a temporal window (typically 4 frames). This design improves compression but introduces a fundamental limitation: a single latent slice does not correspond to a single pixel frame. All compared conditioning strategies differ primarily in how they attempt to recover frame-level alignment from these window-based latents.

1. **Latent-Space Conditioning.** This method encodes the entire groundtruth video with the causal VAE and takes the latent slice whose receptive field overlaps the target frame. However, because each latent mixes information from its surrounding N frames, the "condition latent" inevitably contains content from neighboring frames as well. Thus it does not represent the target frame alone. This explains why in Fig. 7 the PSNR peak does not occur at the correct frame index: the latent window is misaligned with the temporal position of the condition. Furthermore, conditioning the diffusion model on a temporally mixed latent suppresses motion, producing the collapse observed in Dynamic Degree.
2. **Pixel-Space Padding.** This strategy constructs a clip where only the target frame is present and all other frames are zero-padded, then encodes it using the video VAE. Zero-padded frames fall inside the VAE's temporal window, causing the encoder to fuse blank and valid content—an out-of-distribution scenario that leads to color shifts and texture distortion (as shown in Fig. 8). Thus, although this method is temporally precise, its reconstructions are of low fidelity.
3. **Nearest-slot Assignment (w/o RoPE Interpolation).** To avoid temporal mixing, we instead encode each conditional frame independently in image mode, which is spatially robust. However, image-mode latents are continuous in time, whereas the diffusion transformer expects discrete latent slots determined by the VAE stride. Assigning each frame to its nearest temporal slot yields coarse alignment and explains the misaligned PSNR peaks in Fig. 7.
4. **Our Method: Independent Encoding + Temporal RoPE Interpolation.** Our approach resolves all issues above. By encoding each conditional frame independently, we avoid the temporal entanglement intrinsic to video VAEs. Our Temporal RoPE Interpolation then maps each independently encoded latent to an arbitrary continuous temporal coordinate while preserving ordering and positional geometry. This provides precise pixel-frame alignment without relying on video-VAE temporal windows or zero-padded inputs.

C.3. Evaluation Metrics

For completeness, we provide the full set of evaluation metrics used in our quantitative comparisons. In addition to video fidelity measured by FVD [46], we adopt the comprehensive VBench [21] suite, which evaluates both video quality and temporal consistency across multiple dimensions. Specifically, the following metrics are reported:

- **FVD** [46] (↓): Measures overall video fidelity and temporal coherence with respect to real data.
- **Aesthetic Quality** [29] (↑): Assesses the perceptual attractiveness of the generated video frames.
- **Background Consistency** (↑): Evaluates how well the background remains stable across time. This differs from CSCV [4] (better for videos with significant camera motion or large scene transition), which measures adjacent-frame CLIP similarity, whereas VBench computes consistency relative to the first frame.
- **Dynamic Degree** [43](↑): Quantifies the amount of motion present in the generated video, reflecting whether the dynamics are neither overly static nor excessive.
- **Imaging Quality** [25] (↑): Measures sharpness, clarity, and reconstruction of fine-grained details.
- **Motion Smoothness** (↑): Captures the temporal stability of motion across adjacent frames.
- **Overall Consistency** (↑): Evaluates global temporal coherence across the entire clip.
- **Subject Consistency** (↑): Measures identity and appearance stability of the primary subject across time.
- **Temporal Flickering** (↑): Detects flickering artifacts or frame-to-frame instability.
- **Normalized Average** (↑): The mean of all above normalized VBench metrics, providing an aggregated measure of overall video quality.

These metrics jointly offer a detailed and reliable characterization of video quality and temporal stability, ensuring a fair and comprehensive comparison across all evaluated methods.

D. VideoCanvasBench Construction Details

This section provides a comprehensive overview of the data curation and task generation pipeline for *VideoCanvasBench*, the first systematic evaluation suite for arbitrary spatio-temporal video completion.

D.1. Data Curation

We curate two complementary types of sources: (1) *homologous* videos for testing fidelity within a single coherent scene, and (2) *non-homologous* images and videos for evaluating creativity across distinct content.

Homologous Video Set (100 Videos). We began with an initial pool of ∼2,000 videos from Pexels [39]. A multi-stage filtering pipeline was applied to ensure quality and diversity:

- Blur filtering: blurry videos were removed by calculating the CV2.Laplacian [3] score for each frame and excluding those below a threshold of 200.
- Motion filtering: static or nearly-static clips were excluded using RAFT-based motion magnitude thresholds exceeding 5 [43].
- Length filtering: only videos longer than 5 seconds were retained.

From this pool, we selected 100 diverse, high-quality clips covering a wide range of scenes (e.g., human activities, animals, landscapes). All were standardized to 77 frames at 15 FPS to provide a consistent evaluation length. Each video is paired with captions generated by a captioning model fine-tuned on Koala36M [51] following the LLaVA-based [32] annotation pipeline. All captions are further verified by human annotators to ensure accuracy in both content and motion descriptions.

Non-Homologous Image and Video Sets. To test the ability to synthesize across unrelated contexts, we manually curated visually distinct sources from Pexels [39] and Unsplash [45], ensuring large appearance and semantic gaps. The set includes:

- 50 pairs of non-homologous images, selected to maximize dissimilarity (e.g., indoor vs. outdoor, object vs. scene).
- 50 triplets of non-homologous images, further increasing combinatorial diversity.
- 30 pairs of non-homologous video clips, curated for challenging video transitions, similar to the blending function of Sora [37].

These non-homologous cases explicitly test the model's capacity for creative interpolation and cross-scene reasoning. Each non-homologous source is annotated with captions automatically generated by Gemini 2.5 Pro [8] and manually corrected to ensure faithful descriptions of both appearance and motion.

D.2. Benchmark Task Definitions

Task 1: AnyI2V (Any-Timestamp Image-to-Video). This task uses full frames as conditions to test temporal reasoning and interpolation fidelity. We explicitly construct nine sub-tasks by combining conditions from fixed temporal anchors: start (frame 1), middle (frame 41), and end (frame 77).

- Homologous cases. From each source video we sample three anchor frames (start, middle, end), and construct:
  - *Single-frame I2V:* start → video, middle → video, end → video.
  - *Two-frame I2V:* start+end → video, start+middle → video, middle+end → video.
  - *Three-frame I2V:* start+middle+end → video.
- Non-homologous cases. For curated pairs of images, we construct the three two-frame tasks (start+end, start+middle, middle+end). For curated triplets of images, we construct the three-frame task (start+middle+end). Each non-homologous source is annotated with captions automatically generated by Gemini 2.5 Pro [8] and manually checked for accuracy.

Task 2: AnyP2V (Any-Timestamp Patch-to-Video). This variant follows the same nine sub-task definitions as AnyI2V setting, but replaces each full-frame condition with a cropped patch.

- Patch extraction. For each conditional frame, patches are obtained via a semi-automated process: 50% object-aware masks using SAM [26] or YOLO [44], and 50% random crops.
- Temporal anchors. The same start, middle, and end frame positions are used to construct single-, two-, and three-frame variants, for both homologous and non-homologous cases.
- Difficulty. The subset explicitly includes challenging cases with very small subjects, requiring the model to extrapolate from minimal context.

Task 3: AnyV2V (Transition, Inpainting and Outpainting). This task evaluates more general video-level completion scenarios beyond frame- or patch-level control. It consists of three sub-categories:

- *Video Transition.* For 30 curated pairs of non-homologous video clips, the first clip provides the start segment and the second the end segment, while the model synthesizes the intermediate transition. This setup parallels the blending function explored in Sora [37]. Each case is annotated with captions generated by Gemini 2.5 Pro [8] and manually corrected to ensure faithful descriptions of both content and motion.
- *Inpainting.* For homologous videos, interior rectangular masks are applied to each frame, covering 20%–50% of the width/height. The model must fill the missing regions with temporally consistent content.
- *Outpainting.* Boundary masks are applied to crop the central region, masking out 60%–90% of the width/height. The model is required to extrapolate plausible outer regions beyond the visible content.

D.3. Scale

In total, *VideoCanvasBench* includes over **2,000** test cases: 900 for AnyP2V, 900 for AnyI2V, and 230 for AnyV2V. Each case is designed to probe a specific aspect of fidelity, creativity, or temporal reasoning in the proposed unified task.

D.4. Licensing and Annotations.

All videos in our benchmark are sourced from Pexels [39], and images are sourced from both Pexels and Unsplash [45]. Content on Pexels is provided under the Pexels License, which permits free use for commercial and non-commercial purposes without requiring attribution, with restrictions against reselling unaltered copies, use in trademarks, or misuse of identifiable people or brands. A subset of Pexels content is explicitly marked as Creative Commons Zero (CC0), which places the work in the public domain. Unsplash photos are provided under the Unsplash License, which similarly allows free commercial and non-commercial use without attribution, while prohibiting resale of unaltered content, creation of competing stock services, or misleading association with brands or people. In both cases, all curated data is legally licensed for academic research use.

Captions generated by Gemini 2.5 Pro [8] were manually verified by the authors to ensure accuracy and consistency across all benchmark cases.

E. More Analysis and Results

E.1. Analysis of Zero-Padded Inputs

In Section 2, we describe using zero-padding to indicate unconditioned regions when preparing conditional frames. This approach is crucial for our spatial conditioning strategy, as it allows us to precisely specify the location of a condition patch within a frame without modifying the pre-trained VAE backbone. However, a critical question arises: can a standard hybrid video VAE, trained on natural images and videos, effectively handle inputs that contain large areas of zero-valued pixels (i.e., spatial padding)? As illustrated in Figure S10 and Figure S11, this distinction between spatial and temporal padding is fundamental to understanding our method.

To address this, we conducted an empirical study using two popular pre-trained VAE models: Hunyuan I2V and CogVideo. We evaluated their robustness to both spatial and temporal padding under realistic conditions.

Setup. We collected 20 diverse full-resolution images and 20 short video clips from YouTube, representing a wide range of content (e.g., landscapes, cityscapes, indoor scenes, moving vehicles). For each image, we applied random spatial zero-padding masks, covering approximately 40-60% of the pixels. For each video clip, we created three types of padded inputs: 1. A video with conditional frames containing the original content, while all other frames are filled with zeros (pure temporal padding). 2. A video where conditional frames contains cropped region of the original content, with all other frames being zero (temporal & spatial padding).

Each input was then encoded and decoded using the two hybird VAE model. We measured the reconstruction fidelity using PSNR and qualitatively inspected the outputs.

Reconstruction Results. The results provide clear evidence of the differential impact of padding modes:

Spatial Padding Robustness: As shown in Figure S10, both VAE models demonstrate remarkable tolerance to spatial zero-padding. The average PSNR of reconstructed images with spatial padding is only marginally lower than that of the baseline (full image), with an average drop of **0.89 dB**(Hunyuan I2V) and **1.13 dB**(CogVideo).

Temporal Padding Vulnerability: In stark contrast, Figure S11 reveals the limitations of traditional approaches. When applying temporal zero-padding (encoding a single frame into a sequence where most frames are zero), both VAE models exhibit a dramatic degradation in reconstruction quality. The average PSNR drops by over **6.12 dB**(Hunyuan I2V) and **7.01 dB**(CogVideo) compared to the baseline.

Conclusion. These findings confirm that the key to achieving pixel–frame-aware control lies in decoupling spatial and temporal handling. Our method leverages the inherent robustness of the VAE to spatial padding while bypassing the ineffectiveness of temporal padding through our proposed Temporal RoPE Interpolation. This separation enables flexible, high-fidelity video completion using a frozen VAE without requiring retraining or architectural modification.

In addition to the controlled analyses presented above, our qualitative results under arbitrary spatiotemporal conditioning (Fig. 1 and Fig. S14) further demonstrate that spatial zero-padding remains stable even under large variations in placement, content, and temporal context. These observations provide complementary evidence supporting the effectiveness and general applicability of our approach across diverse zero-padded settings. Together with the quantitative results reported in the main text, these findings consistently validate the necessity and effectiveness of our design.

E.2. Advantages of Temporal RoPE Interpolation

Figure 7 in the main paper has shown that our Temporal RoPE Interpolation achieves *precise one-to-one alignment* between condition frames and their target temporal positions. Here we further demonstrate not only that our model can leverage this precision for *dense* conditioning, but also why this capability represents a crucial advantage over competing paradigms.

To this end, we conduct an additional experiment on the homologous video set from *VideoCanvasBench*. Each 77-frame video is conditioned on the first five frames (0-4) in two different ways:

- **Sparse Condition:** Only the boundary frames (0 and 4) are provided. The model must interpolate the three missing frames (1, 2, 3) in between.

Figure S10. **Robustness of Hybrid Video VAEs to Spatial Padding**. This figure demonstrates that both the Hunyuan I2V and CogVideo VAE models can tolerate spatial zero-padding well. When reconstructing images with large zero-padded regions (middle row), the PSNR values are only slightly lower than those of the full, unpadded images (top row). Crucially, the original content within the non-zero regions is faithfully preserved, while the padded areas remain visually neutral. This empirical evidence confirms that our spatial conditioning strategy, which relies on zero-padding before VAE encoding, is stable and practical, enabling precise spatial control without degrading the quality of the conditioned content.

- **Dense Condition:** All five frames (0, 1, 2, 3, 4) are explicitly provided as conditions, testing frame-wise alignment at every step.

Both settings are used to generate the full video, and we evaluate the fidelity by computing PSNR on the first 5 frames against the ground truth.

The quantitative results in Table R4 confirm that explicitly conditioning on consecutive frames yields higher reconstruc-

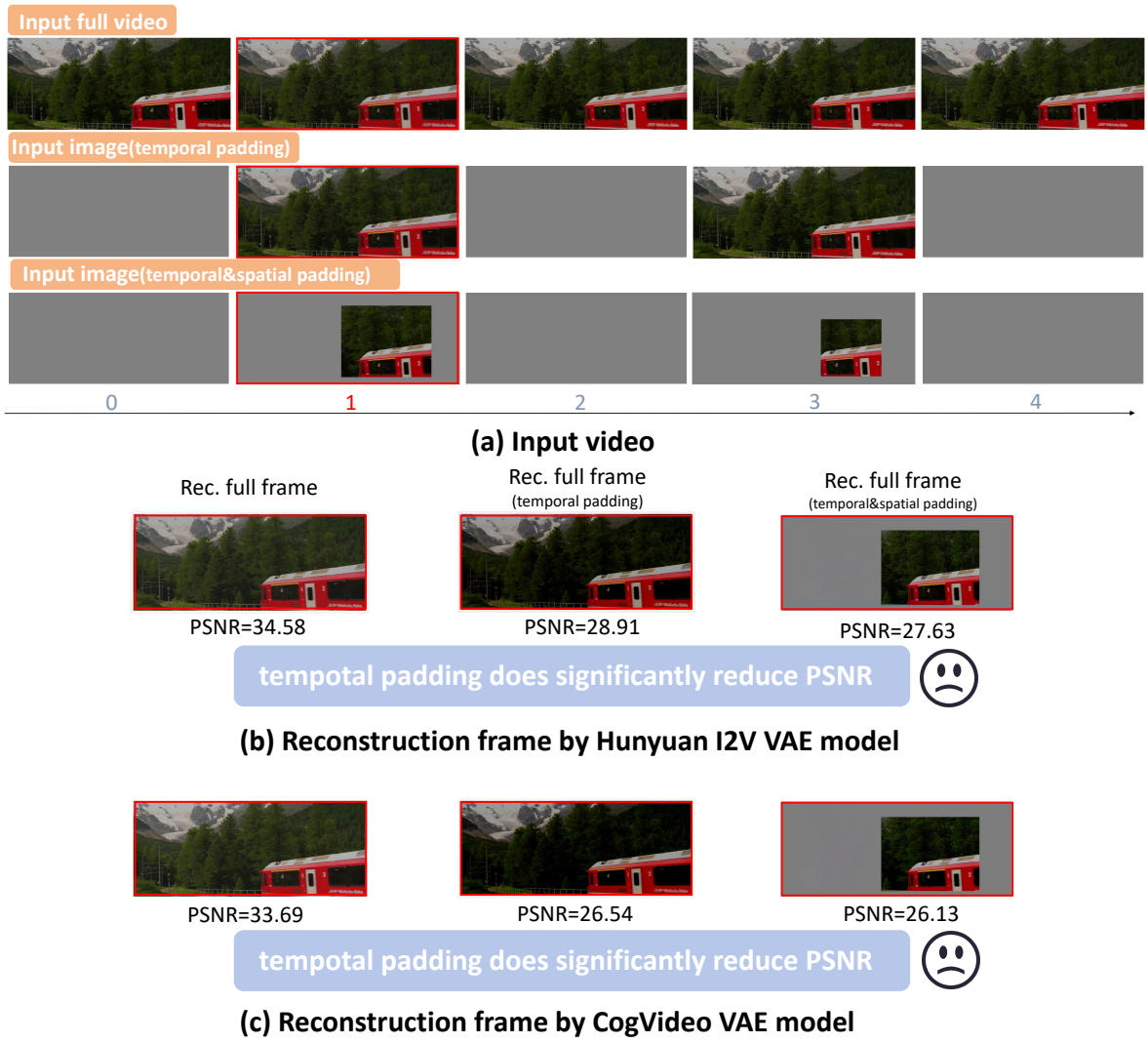


Figure S11. **Vulnerability of Hybrid Video VAEs to Temporal Padding**. This figure contrasts the robustness observed in spatial padding. When applying temporal zero-padding (where only specific frames contain content), both VAE models suffer a relatively great drop in reconstruction quality. The PSNR values for the padded reconstructions (bottom rows) are much lower than those of the full video (top row), demonstrating a degradation in fidelity. The reconstructed frames exhibit noticeable color shifts, and loss of detail, highlighting that the VAE cannot handle such distributionally mismatched inputs. This mode underscores why direct temporal zero-padding is ineffective and validates the necessity of our Temporal RoPE Interpolation strategy, which avoids this problem by operating at the latent token level with fractional positions.

Table R4. Average PSNR (dB) across 100 videos under sparse vs. dense conditioning.

Condition Type	Conditioned Frames	PSNR (↑)
Sparse (two frames)	0, 4	24.789
Dense (five frames)	0, 1, 2, 3, 4	**25.033**

tion fidelity. Figure S12 provides a visual illustration. In the sparse case, our model generates a plausible interpolation, but with minor, expected drift in the unconditioned intermediate frames. The dense case, in contrast, achieves a near-perfect reconstruction.

This comparison highlights a fundamental limitation of paradigms like *Channel Concatenation*. Due to their coarse, slot-based nature and the constraint of a frozen VAE, they can only condition on one frame per latent slot (e.g., one frame for every $N = 4$ pixel frames). They are therefore structurally incapable of providing dense guidance for the intermediate frames (e.g., frames 1, 2, 3) and are locked into a "sparse" conditioning mode, inevitably suffering from the

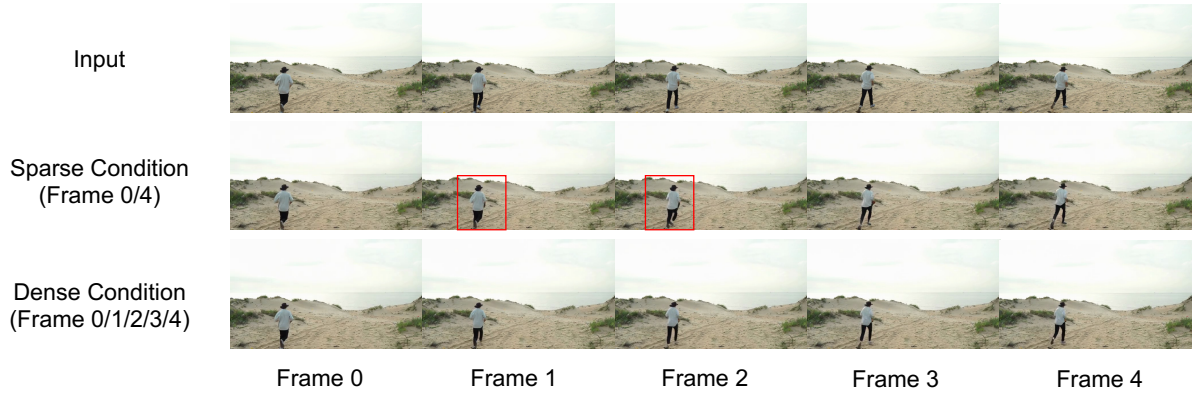


Figure S12. **Visual comparison of Sparse vs. Dense conditioning.** The top row shows the ground-truth frames. The middle row (**Sparse**) is generated using only frames 0 and 4 as conditions; note the plausible but slightly drifted interpolation for the intermediate frames. The bottom row (**Dense**) is generated using all five frames as conditions, resulting in a near-perfect reconstruction. This highlights the benefit of dense, frame-by-frame control—a capability unique to our method.

kind of interpolation drift shown in our sparse example. In contrast, our *Temporal RoPE Interpolation* uniquely enables true dense conditioning, allowing VideoCanvas to maintain high fidelity frame-by-frame—a capability that is structurally inaccessible to these competing methods.

E.3. More Ablation and User Studies

E.3.1. Ablation of RoPE Strategy

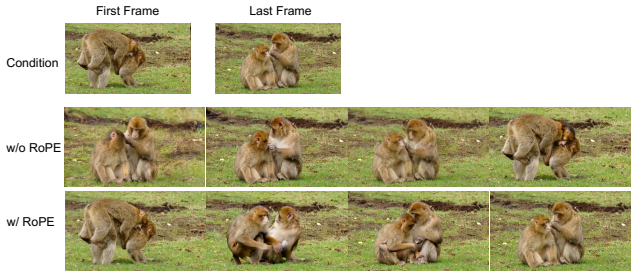


Figure S13. Visualization of without RoPE.

We further investigate the impact of different Rotary Positional Embedding (RoPE) variant strategies on temporal alignment. Specifically, we compare our proposed **Fractional Interpolation** (mapping frames to $0, 0.25, \ldots, 19$) against an **Integer Extrapolation** strategy (scaling indices to $0, 1, \ldots, 76$ for a $4\times$ expansion) and a baseline **w/o RoPE**.

As reported in Table R5, our method achieves the best performance with the lowest FVD score. We attribute the superiority of our fractional strategy to the preservation of pre-trained priors. Since the base model was trained on a specific temporal range (indices $0 \sim 19$), our fractional approach ensures the input indices remain within this learned distribution. This allows the model to effectively inherit the original temporal priors, leading to faster convergence and better performance compared to the Integer method, which shifts the distribution to an unfamiliar range ($0 \sim 76$).

Table R5. **Quantitative ablation of RoPE strategies.** We report the FVD (↓) scores (lower is better) on the validation set. Our fractional interpolation strategy outperforms both integer extrapolation and the baseline without RoPE.

Method	Index Mapping Strategy	FVD (↓)
w/o RoPE	N/A	12.568
Integer Extrapolation	$0, 1, \ldots, 76$	11.079
Ours (Fractional)	$\mathbf{0, 0.25, \ldots, 19}$	**10.943**

Furthermore, removing RoPE proves disastrous. Without explicit positional indicators, the condition tokens injected via concatenation lack the temporal cues necessary to "lock" onto specific frame positions. Consequently, the mechanism degenerates from precise conditioning into a loose "image reference" mode, where the generated frames fail to strictly adhere to the condition frames. This degradation is visually evident in Fig. S13, where the model fails to maintain temporal position consistency.

E.3.2. User Study on Conditioning Paradigms

To complement the quantitative results in Table 2, where our method achieves the best FVD and Dynamic Degree, we further conducted a human preference study using the diverse tasks defined in the VideoCanvas benchmark. We compare our **In-Context Conditioning (ICC)** against the two representative paradigms: *Latent Replacement* and *Channel Concatenation*.

Setup. We sampled generated videos across the broad range of tasks in VideoCanvas. Evaluators performed pairwise comparisons based on visual quality, motion smoothness, and condition fidelity.

Results. As shown in Table R6, our method consistently outperforms both baselines:

- **Vs. Latent Replacement:** Ours achieves a **65.4%** win rate. Users penalized the replacement method for its lower dynamic degree (consistent with Table 2) and disjointed motion transitions.
- **Vs. Channel Concatenation:** Ours wins by **48.2%** (with a significant tie rate). This confirms that our parameter-free ICC strategy is more effective than increasing channel dimensionality (16.6M extra params), yielding better visual harmony.

These results align with the quantitative metrics in Table 2, verifying that ICC provides the most effective guidance for the diffusion model.

Table R6. **Human Preference Evaluation.** Comparison of our In-Context Conditioning (Ours) against two representative paradigms: *Latent Replacement* and *Channel Concatenation*. We report the percentage of user votes favoring our model versus the variants, averaged across all evaluated tasks on the VideoCanvas benchmark.

Comparison	Preference Rate (Ours vs. Variant)		
	Win (↑)	Tie	Loss (↓)
Ours vs. Latent Replacement	**65.4%**	22.1%	12.5%
Ours vs. Channel Concatenation	**48.2%**	28.3%	23.5%

E.4. Training and Inference Cost

Table R7. Training and inference cost comparison across paradigms. Training time is measured over 20k steps. Inference time is per 77-frame video at 384×672 with different numbers of conditional frames.

Method	New Params	Train	Inference		
			1 frame	2 frame	3 frame
Replacement	0	21.47h	159s	159s	159s
Channel Concat	**16.6M**	22.47h	164s	164s	164s
ICC (Ours)	0	24.54h	168s	175s	184s

Tab. R7 compares the computational cost of different conditioning paradigms. Unlike channel concatenation, which relies on a 16.6M projection layer, our ICC design introduces no additional parameters, and the training cost remains comparable (24.5h vs. 21–22h) since ICC only adds lightweight spatio-temporal tokens.

During inference, ICC exhibits a content-aware and controllable scaling: the compute grows with the number of conditioning frames, because a richer conditioning context requires processing longer input sequences within the transformer. For a 77-frame video at 384×672, inference takes 168 s with a single conditioning frame, and gradually increases to 333 s / 520 s / 910 s when conditioning on 20 / 40 / 77 frames, respectively.

In sparse conditioning scenarios, the additional cost is almost negligible, since only a few frames expand the sequence length. When more frames are provided, ICC allows users to intentionally trade additional computation for higher fidelity on the corresponding conditioned timestamps, offering stronger control instead of incurring unavoidable overhead. While this scaling makes ICC marginally slower than baselines with fixed-cost inference, the trade-off is justified, as ICC consistently yields higher fidelity and better spatio-temporal alignment (see Sec. 4.2, Tab. 1 and Tab. 2).

F. Applications and Qualitative Results

F.1. Applications

The teaser figure (Fig. 1) has shown some cases, and in this section, we provide extensive qualitative results to demonstrate the versatility and effectiveness of our VideoCanvas framework across a wide range of applications.

- **Any-Timestamp Patch-to-Video (AnyP2V).** In Figure S14, we demonstrate our core capability of generating a complete video from a varying number of sparse patches. We showcase challenging scenarios using one, two, three, and even four conditional patches, placed at arbitrary timestamps to rigorously test the model's spatio-temporal reasoning beyond simple first-frame conditioning.
- **Any-Timestamp Image-to-Video (AnyI2V).** Figure S15 illustrates the flexibility of our framework on full-frame conditions. The examples include standard cases like first-frame I2V and first-last-frame interpolation, as well as more challenging scenarios where conditions are placed at arbitrary middle timestamps, a capability not well supported by prior methods.
- **Video-Level Completion and Creation (AnyV2V).** Our framework naturally unifies a variety of video editing and creation tasks within a single model. We provide examples of:
  - **Video Transition:** Creative transitions between non-homologous clips are demonstrated in Figure S16.
  - **Video Painting:** Inpainting and outpainting results are shown in Figure S17, where the red dashed contours indicate the generated regions.
  - **Video Extension and Looping:** As demonstrated in Figure S18, we showcase long-duration synthesis by extending short clips to over a minute in length while maintaining temporal consistency. This capability can be guided by interactive text prompts to evolve the narrative. Furthermore, we can create perfectly seamless loops by generating a smooth transition from the video's end back to its beginning. Our approach leverages motion context from the last segment's frames to effectively avoid the stuttering artifacts that are common in naive first-last frame-looping methods.

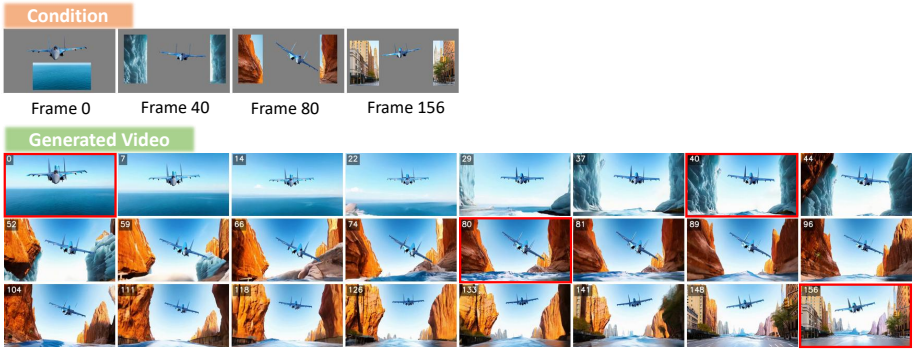
– **Video Camera Control:** As demonstrated in Figure S19, our framework can emulate camera cinematography by progressively translating or scaling content on the spatio-temporal canvas. This enables a variety of standard camera effects, such as zooms and pans. We showcase this capability by applying dynamic camera movements to classic movie shots, demonstrating its potential for creative post-production.

F.2. Qualitative Comparison

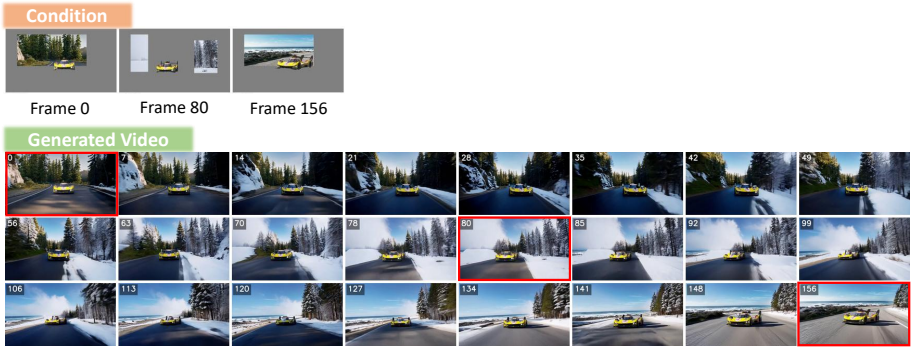
The following figures showcase side-by-side comparisons with baseline paradigms, illustrating our method's superior performance in motion smoothness, detail sharpness, and temporal consistency.

Fig. S20 and Fig. S21 present qualitative results across the six varied tasks (I2V, FLF2V, TF2V, FLP2V, Inpainting, and Outpainting), visually confirming our framework's strong and consistent performance across these diverse domains.

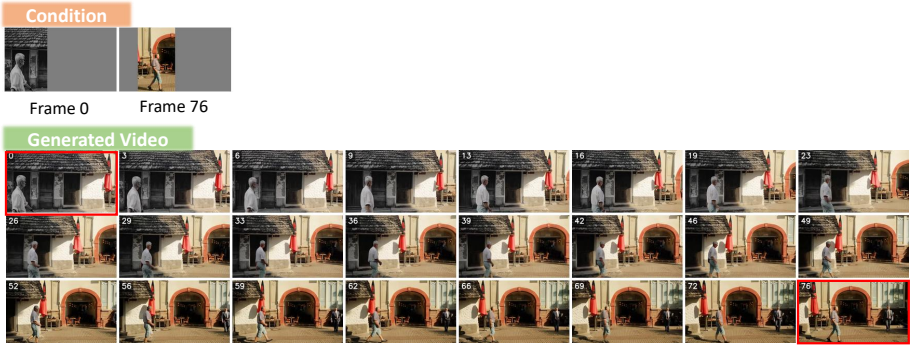
Finally, Figure S22 provides additional direct comparisons against different paradigms across a diverse set of challenging cases, further highlighting the robustness and superiority of our approach.



(a) **case 1** **“**Blue airplane races ocean, glaciers, canyon, into bustling city…”



(b) **case 2** **“**Yellow race car speeds forest, snow, then coastal highway...”

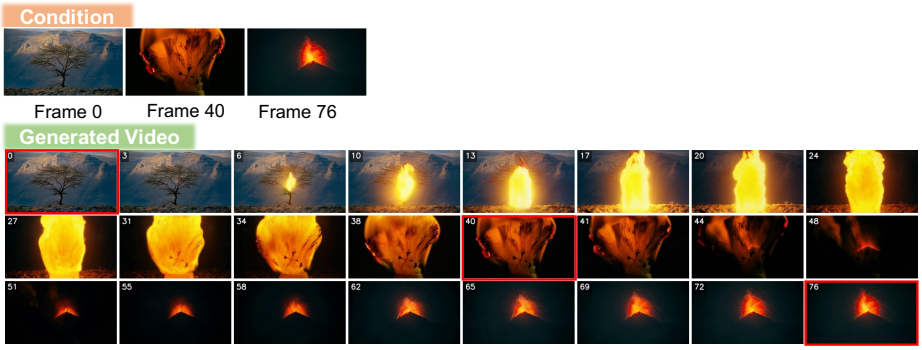


(c) **case 3** **“**Elderly man, old house morphs into modern café...”

(d) **case 4** **“**Young couple with dog walking serene autumn path…”

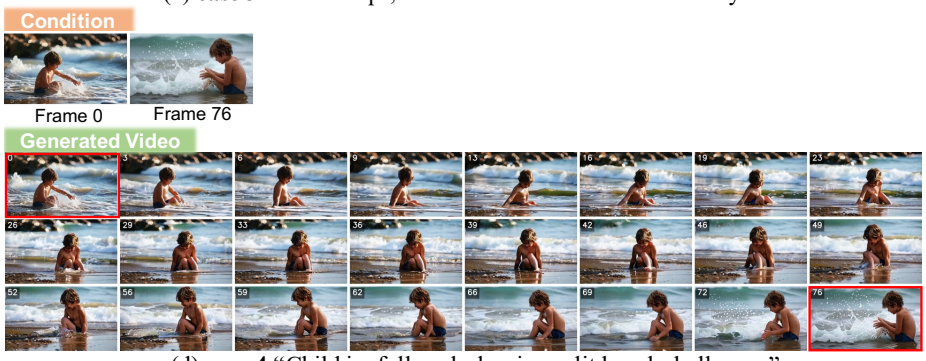
Figure S14. **Results on Any-timestamp Patches to Videos.**

(a) **case 1** **"**Young woman holds purple flowers by sunlit creek..."



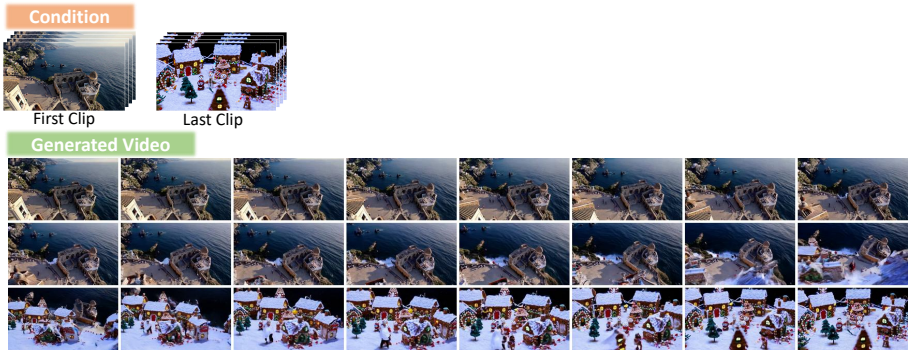
(b) **case 2** **"**Tree morphs into glowing object, erupts as volcano..."

(c) **case 3** **"**Zebra leaps, transforms into colorful kite in sky..."

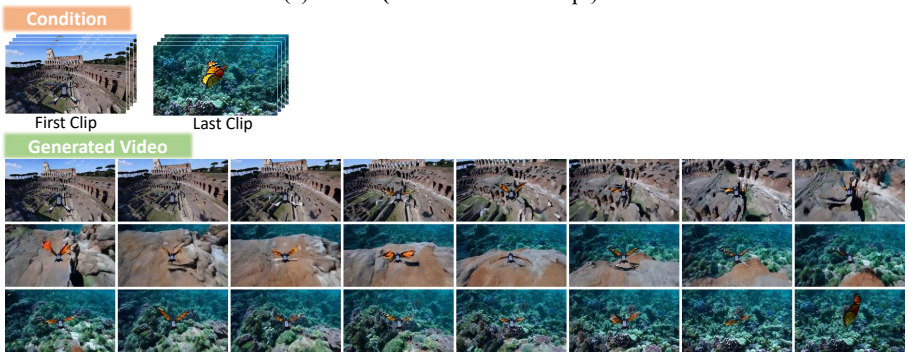


(d) **case 4** **"**Child joyfully splashes in sunlit beach shallows..."

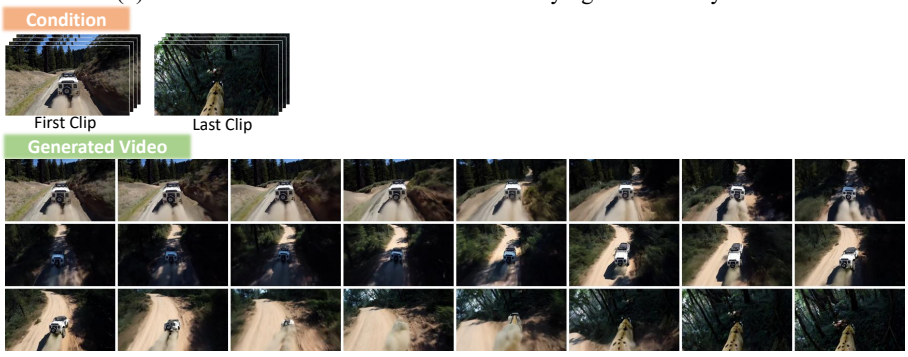
Figure S15. **Results on Any-timestamp Images to Videos.**



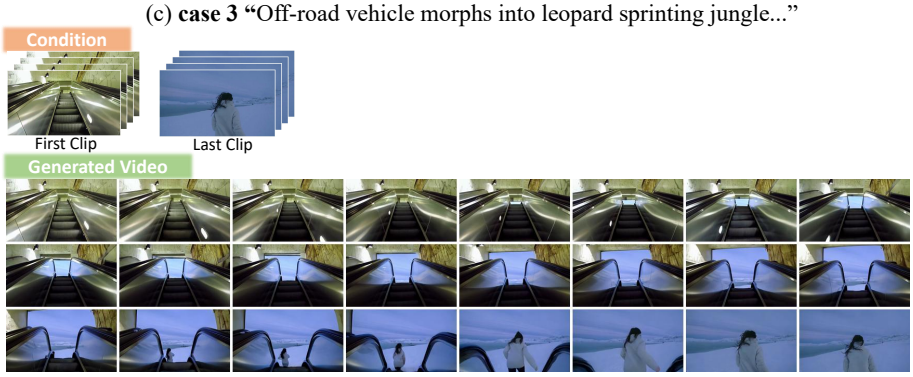
(a) **case 1 (**Without Text Prompt)



(b) **case 2 "**A continuous shot from a drone flying to a butterfly"



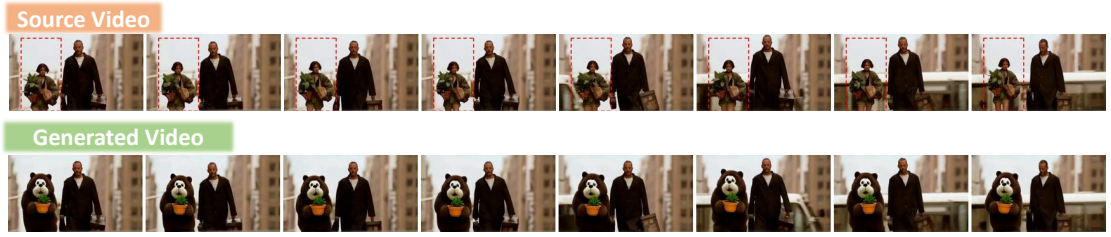
(c) **case 3 "**Off-road vehicle morphs into leopard sprinting jungle..."



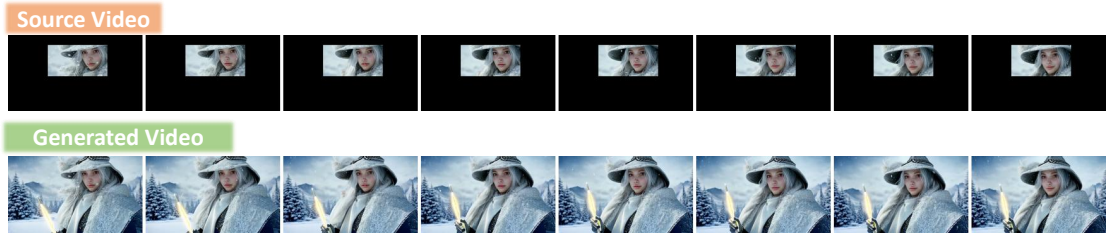
(d) **case 4 "**Escalator ascends into arctic sky, woman runs snowy landscape..."

Figure S16. **Results on Video Transition.**

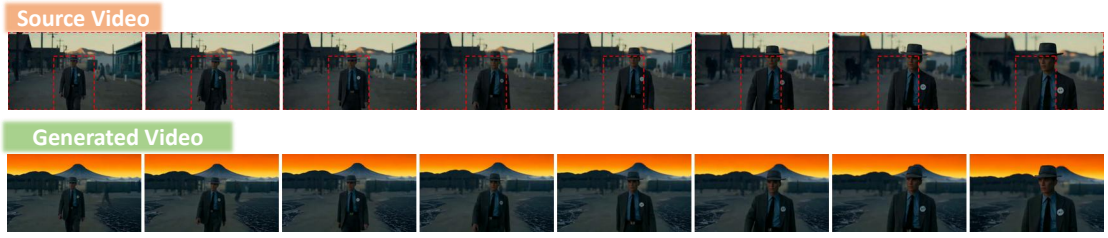
(a) **Video Inpainting case 1 "**Cartoon bear surrenders humorously to rooftop gunman…"



(b) **Video Inpainting case 2 "**Léon and cartoon bear walk, carrying plants together…"



(c) **Video Outpainting case 1 "**Silver-haired sorceress stands in snowy fantasy landscape.…"



(d) **Video Outpainting case 2 "**Volcanic plain with dormant volcano under orange sky.…."

Figure S17. **Results on Video Inpainting and Outpainting.** The red dashed contours indicate the regions that are subject to inpainting or outpainting.

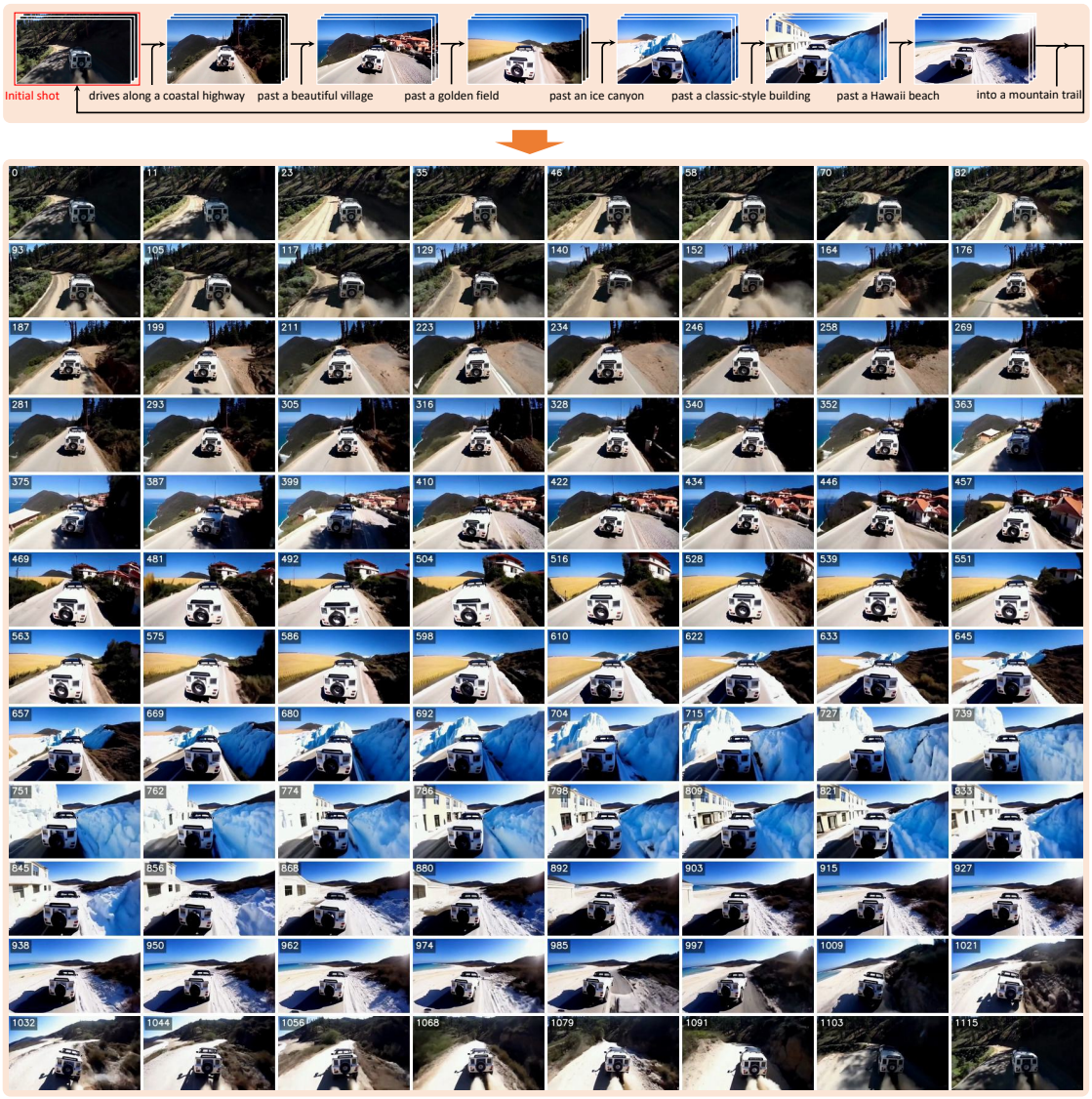
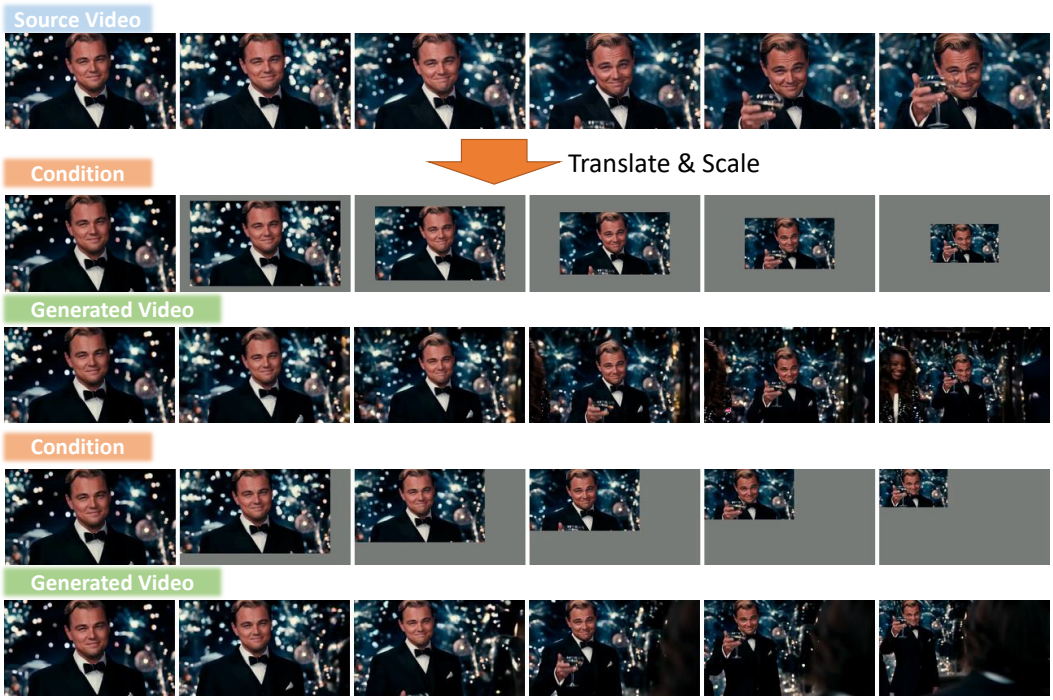
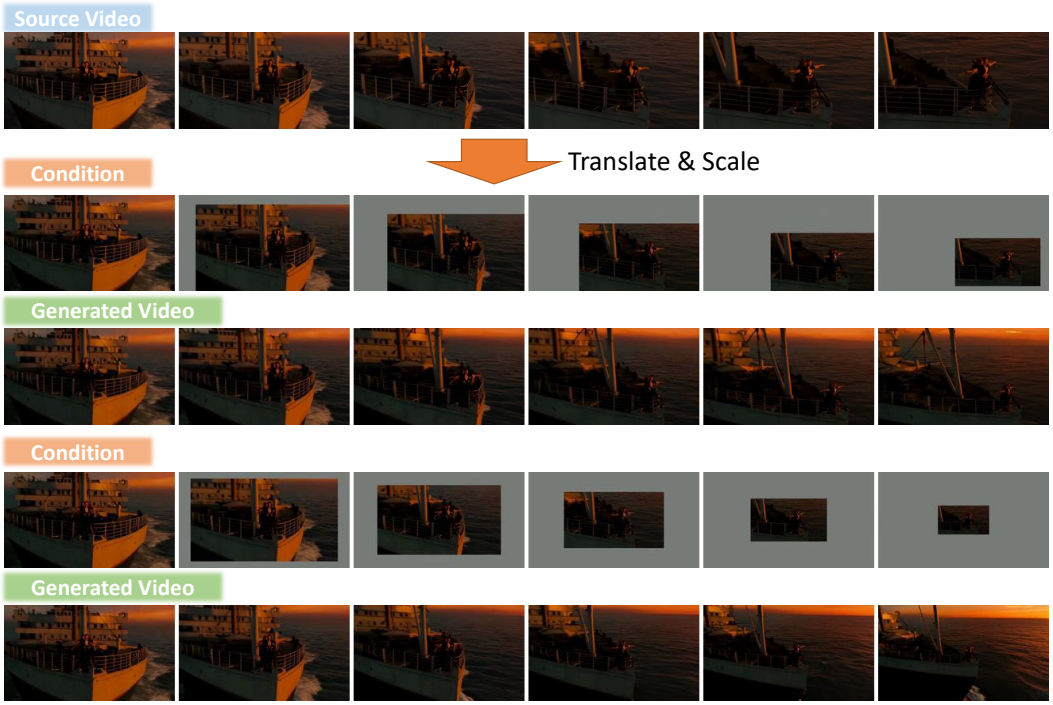


Figure S18. **Results on Video Extension and Seamless Looping.** The example showcases a video extended to **over 1,000 frames** by first applying our video extension capability and then generating a seamless transition back to the initial state. This highlights our model's ability to maintain temporal consistency and visual quality over a long generation horizon without suffering from quality degradation or motion collapse.



(a) Case 1: Video Clip From *The Great Gatsby*



(b) Case 2: Video Clip From *Titanic*

Figure S19. **Results on Video Camera Control.** The examples showcase emulated camera effects such as zoom and pan, achieved by progressively translating and scaling content on the spatio-temporal canvas.

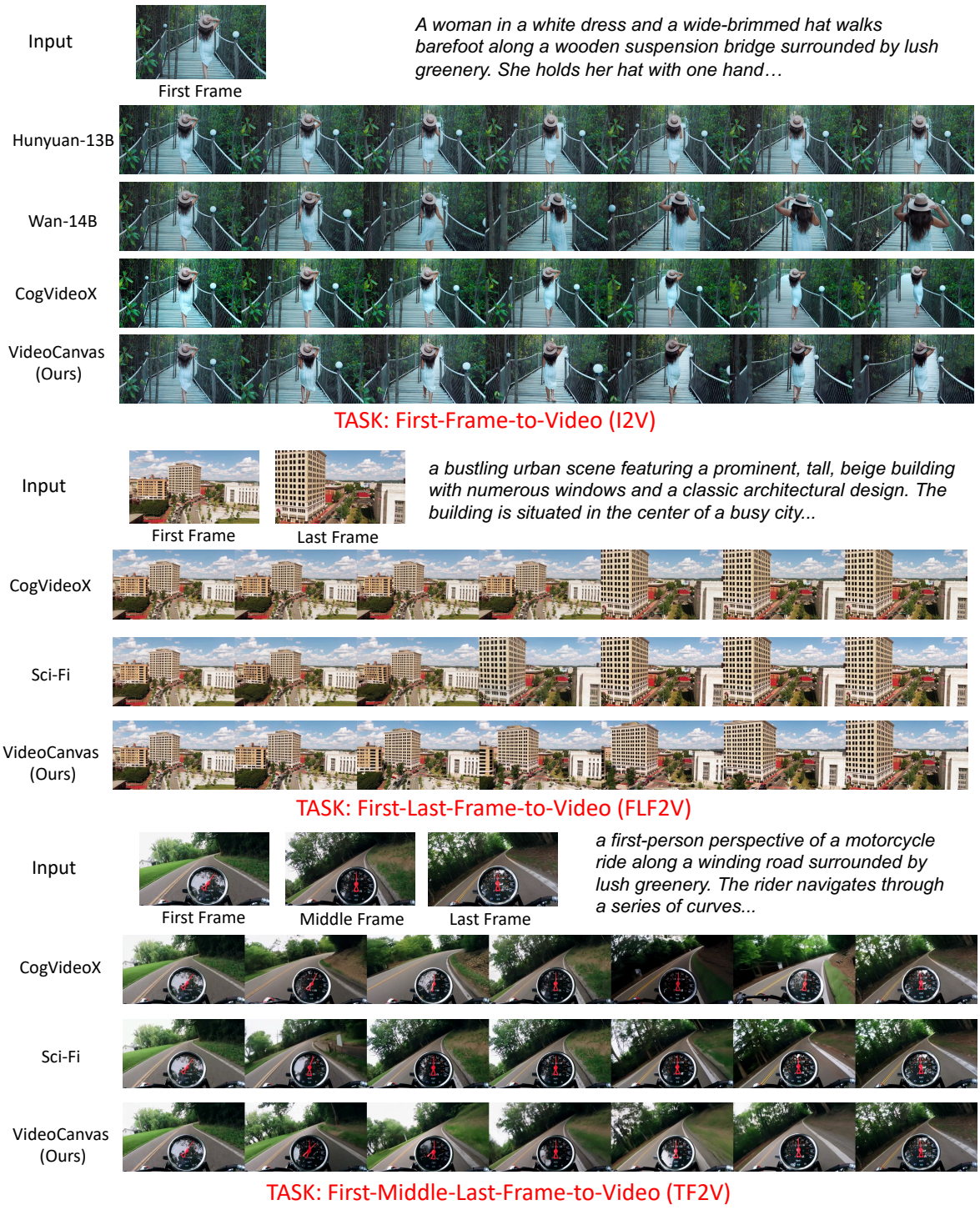


Figure S20. **Comparisons with baseline models (1/2).**

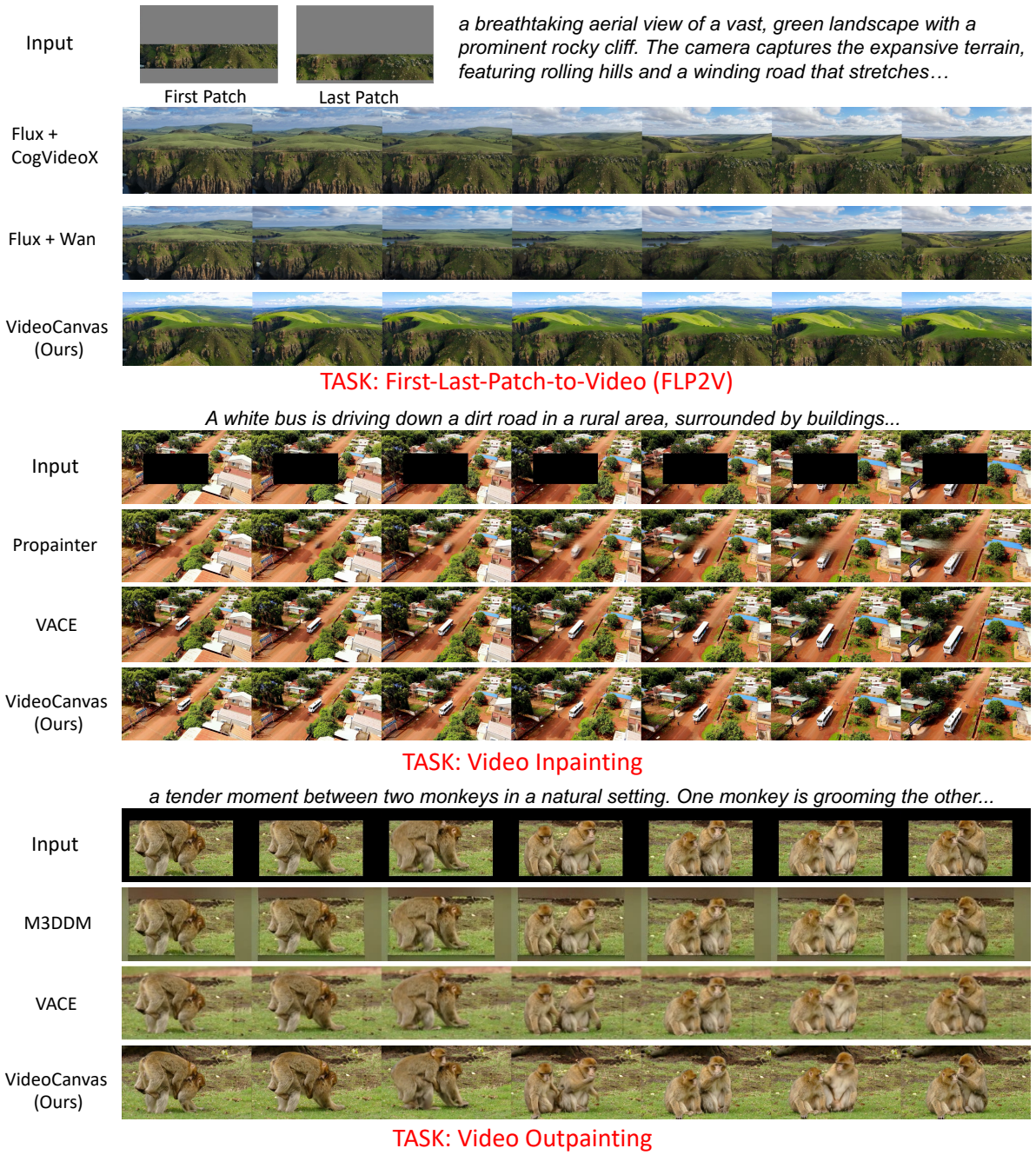
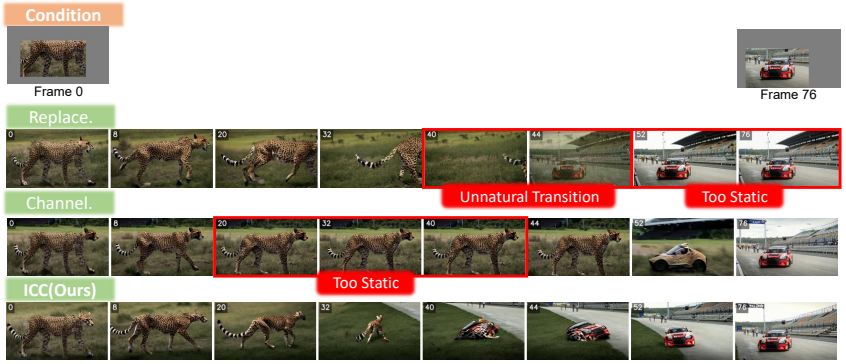
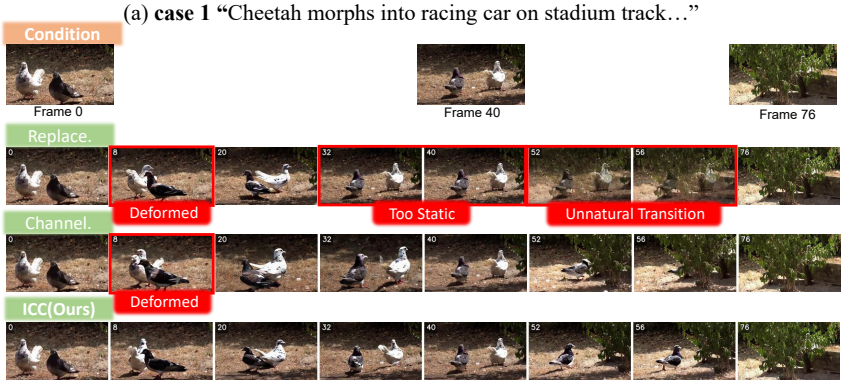


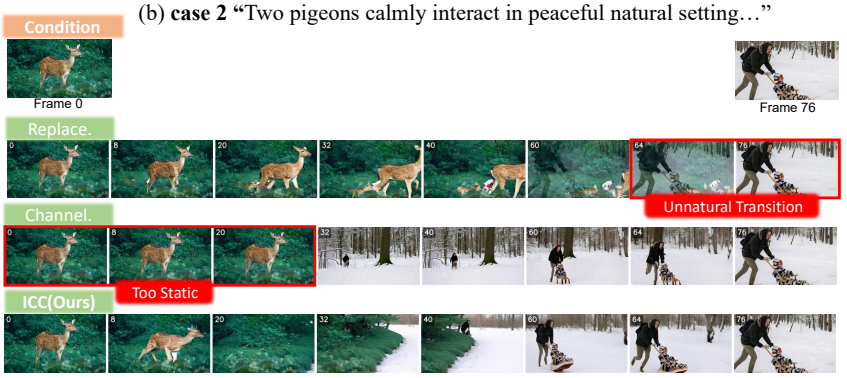
Figure S21. **Comparisons with baseline models (2/2).**



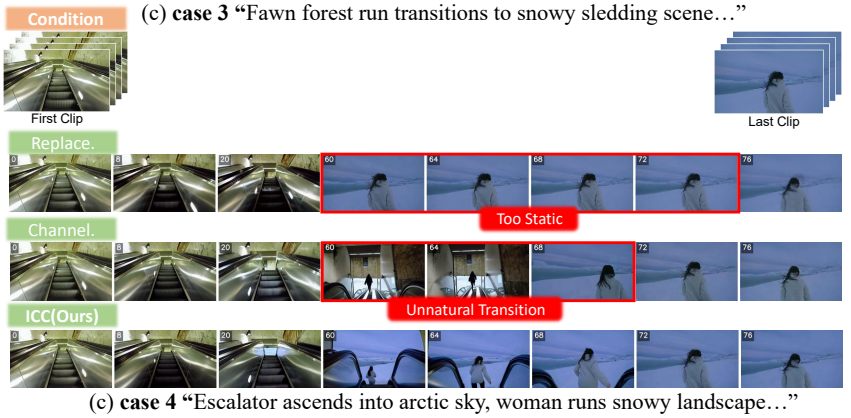
(a) **case 1** **"**Cheetah morphs into racing car on stadium track…"



(b) **case 2** **"**Two pigeons calmly interact in peaceful natural setting…"



(c) **case 3** **"**Fawn forest run transitions to snowy sledding scene…"



(c) **case 4** **"**Escalator ascends into arctic sky, woman runs snowy landscape…"

Figure S22. **Comparisons with baseline paradigms.**