

## 360 A Technical Appendices and Supplementary Material

### 361 A.1 Training Details

362 All models are fine-tuned using an NVIDIA H100 GPU. Our method builds on the CogVideoX-  
363 5B backbone and is fine-tuned with LoRA (rank 128), resulting in approximately 130M trainable  
364 parameters. Training with 49 frames requires roughly 30GB of GPU memory. For ControlNet,  
365 we apply LoRA with the same rank, yielding a comparable parameter count of around 150M, and  
366 requiring approximately 60GB of GPU memory. For Fun-pose, we use the official full fine-tuning  
367 setup, which consumes around 75GB of GPU memory.

### 368 A.2 Training Amount vs. Performance

369 This section demonstrates the training efficiency of our method compared to ControlNet. Figure 5  
370 presents performance curves for various metrics—including CLIP-T, CLIP-I, SSIM, DINO, and  
371 LPIPS—plotted against training time. Our method consistently outperforms ControlNet across all  
372 metrics at equivalent training durations. Moreover, with the exception of CLIP-T, all metrics show a  
373 clear upward trend, indicating continued improvement with more training. In contrast, ControlNet  
374 exhibits no such trend, suggesting that its training style tends to overfit and struggles to generalize  
375 under limited data regimes.

### 376 A.3 Ablation Study

377 We conduct ablation study on various buffer frame designs. Specifically, we compare our default  
378 setting—using a uniformly increasing noise schedule—with alternative strategies: (1) a constant  
379 noise level  $t$  for all buffer frames (denoted as Constant- $t$ , where  $T = 100$ ), and (2) linear-quadratic  
380 schedules with concave or convex profiles. Figure 6 presents both zero-shot and fine-tuned results for  
381 these configurations. While all variants produce reasonable target frames, we observe that the convex  
382 schedule and the constant-25 baseline exhibit poor condition alignment and noticeable artifacts in  
383 the zero-shot setting. After fine-tuning, all methods perform comparably, though our default setting  
384 with uniformly increasing noise remains preferred. Quantitative results after training are presented in  
385 Table 3 and Table 4 for the I2V and V2V tasks, respectively.

386 We also evaluate the effect of varying the number of buffer frames, ranging from 1 to 5, denoted as  
387 Buffer- $n$  in Figure 7. In the zero-shot setting, we observe that all configurations perform comparably  
388 overall; however, shorter buffers tend to produce noisier transitions, likely due to abrupt scene  
389 changes. Conversely, longer buffers show a tendency to weaken the influence of the condition. After  
390 fine-tuning, all variants produce similarly high-quality results.

### 391 A.4 Dataset

392 For the object-to-motion task, we use the DTU dataset [13]. For character-to-video, keyframe  
393 interpolation, and ad video generation tasks, we manually collected condition–video pairs tailored  
394 to each task. For action transfer, we curate videos from SSv2 [7]. In the video style transfer task,  
395 we first synthesize starting frames using FLUX.1-dev [15], and then generate paired videos using  
396 SoRA [18] and Wan2.1 [24]. Each task is trained on 30 samples. All videos contain 49 frames at 10  
397 frames per second (fps), resized to either 480×480 or 848×480 while preserving the original aspect  
398 ratio.

399 For evaluation and demonstration, we use image and video conditions that are not part of the training  
400 set. These include both manually collected images and synthesized ones generated using GPT-  
401 4o, FLUX, and Sora. For the action transfer task, we use unseen video samples from SSv2 [7].  
402 Quantitative evaluations are conducted on 100 samples. For image-based metrics such as CLIP and  
403 LPIPS, scores are computed on a per-frame basis and then averaged to obtain the final results.

404 Training and evaluation prompts are generated using GPT-4o. Each prompt is structured to encompass  
405 the condition, buffer, and target frames, with condition and buffer frames denoted as [CONDITION]  
406 and target frames as [VIDEO]. Below is the full prompt used for the sample in the ablation study:

### TIC-FT prompt

*This animated clip demonstrates the transformation of a static character illustration into a lively and expressive animated figure; [CONDITION] the condition image showcases a cheerful cartoon-style young buffalo with thick brown fur, curved yellow horns, and a big, friendly smile. The character's wide eyes and upright posture are set against a warm orange background, giving it a lively and playful presence. [VIDEO] the video animates the buffalo inside a grand museum, where it wears a red t-shirt and points excitedly at a large dinosaur skeleton behind glass. Its eyes are wide with curiosity and its mouth open in awe, while elegant stone columns and soft lighting emphasize the sense of wonder and fascination with history.*

407

## 408 A.5 Task Descriptions

409 We detail the construction of data and latent sequences for each conditional video generation task used  
 410 in our experiments. All tasks are configured with a total of 13 latent frames, corresponding to 49 video  
 411 frames. While this number can be adjusted based on application needs, we adopt the 13-frame setting  
 412 throughout for implementation simplicity and consistency. The initial latent sequence comprises  
 413 condition frames, intermediate buffer frames, and noised target frames. An exception is the action  
 414 transfer task, where buffer frames are omitted, as the last condition frame serves as the starting frame  
 415 of the target sequence. The specific configurations for each task are described below.

416 **Image-to-Video** This task aims to generate a full video conditioned on a single image. The video  
 417 need not begin directly from the image's visual content; instead, the image may represent a high-level  
 418 concept such as a character profile or an object viewed from the top, with the video depicting novel  
 419 dynamics (e.g., a rotating 360° view).

420 A single reference image is replicated to occupy the first 4 latent frames, followed by 9 target frames.

- 421 • Clean condition: 1 frame (from the image)
- 422 • Buffer: 3 frames (noised condition)
- 423 • Target: 9 frames (pure noise)

424 We visualize the initial latent frames and their denoising process in Figure 8.

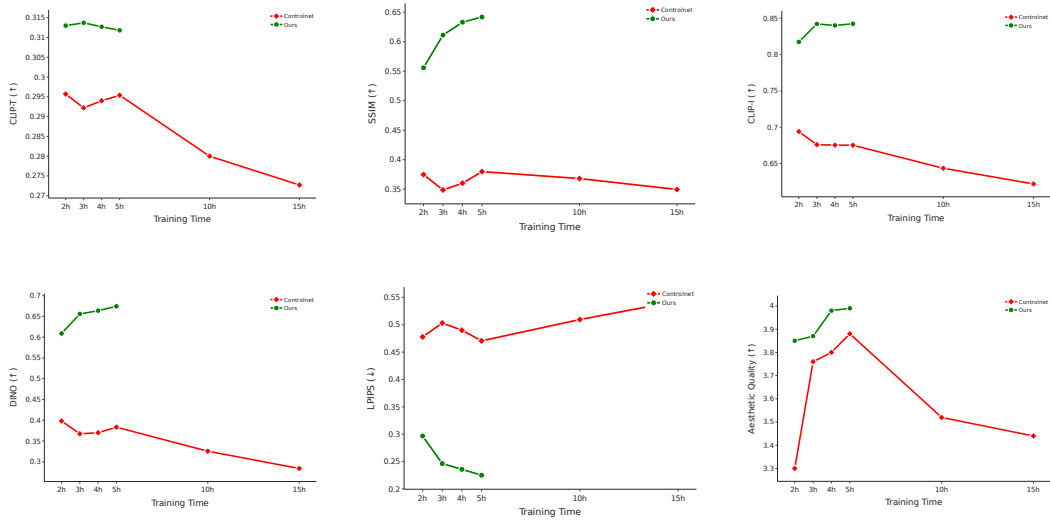


Figure 5: Performance curves for CLIP-T, CLIP-I, SSIM, DINO, and LPIPS metrics plotted against training time. Our method consistently outperforms ControlNet across all metrics at equivalent training durations.

Table 3: Ablation study of constant noise scheduling for buffer frames, evaluated on I2V tasks using VBench, GPT-4o, and perceptual/similarity metrics.

Method	VBench			GPT-4o			Perceptual similarity				
	<i>subject consistency</i>	<i>background consistency</i>	<i>motion smoothness</i>	<i>aesthetic quality</i>	<i>structural similarity</i>	<i>semantic similarity</i>	CLIP-I	CLIP-T	LPIPS↓	SSIM	DINO
Ours	<b>0.9672</b>	0.9729	<b>0.9930</b>	<b>4.13</b>	<b>3.14</b>	3.63	<b>0.8329</b>	<b>0.3143</b>	<b>0.4332</b>	<b>0.5917</b>	<b>0.5530</b>
Constant-25	0.9516	0.9724	0.9920	4.09	2.81	3.45	0.7734	0.3062	0.6088	0.4240	0.4202
Constant-50	0.9509	<b>0.9740</b>	0.9915	4.05	3.01	3.51	0.7760	0.3010	0.6157	0.4188	0.4228
Constant-75	0.9511	0.9722	0.9917	4.02	3.07	<b>3.68</b>	0.7725	0.3003	0.6148	0.4250	0.4259

Table 4: Ablation study of constant noise scheduling for buffer frames, evaluated on V2V tasks using VBench, GPT-4o, and perceptual/similarity metrics.

Method	VBench			GPT-4o			Perceptual similarity				
	<i>subject consistency</i>	<i>background consistency</i>	<i>motion smoothness</i>	<i>aesthetic quality</i>	<i>structural similarity</i>	<i>semantic similarity</i>	CLIP-I	CLIP-T	LPIPS↓	SSIM	DINO
Ours	<b>0.9736</b>	<b>0.9743</b>	<b>0.9935</b>	<b>3.99</b>	<b>3.90</b>	<b>4.41</b>	<b>0.8794</b>	0.3080	<b>0.2298</b>	<b>0.6541</b>	<b>0.6596</b>
Constant-25	0.9539	0.9652	0.9873	3.90	3.55	4.20	0.8037	0.3103	0.2744	0.5785	0.6083
Constant-50	0.9524	0.9652	0.9886	3.88	3.86	4.31	0.8460	<b>0.3153</b>	0.2364	0.6039	0.6528
Constant-75	0.9327	0.9552	0.9821	3.69	3.60	4.25	0.8330	0.3142	0.2797	0.5707	0.6368

**Video Style Transfer** This video-to-video task transforms the visual style of a source video into that of a target domain (e.g., converting a realistic video into an animated version) while preserving motion and structure.

The first 7 latent frames are taken from a source video and the remaining 6 from a style-transferred version.

- Clean condition: 4 frames (from the source video)
- Buffer: 3 frames (noised condition)
- Target: 6 frames (pure noise)

We visualize the initial latent frames and their denoising process in Figure 9.

**In-Context Action Transfer** This task generates a video that continues a novel scene using motion inferred from a source video. Given a reference action and the first frame of a new environment, the model synthesizes future frames that imitate the observed motion within the new context.

The first 6 latent frames are from a reference action video, the 7th is the first frame of a novel scene, and the rest are the continuation.

- Clean condition: 6 frames (from the reference action video)
- Query frame: 1 clean frame (from the novel scene)
- Target: 6 frames (pure noise)

*No buffer frames are used in this task, as the first frame of the target video is explicitly provided as part of the condition.* We visualize the initial latent frames and their denoising process in Figure 10.

**Keyframe Interpolation** This task fills in intermediate frames between sparse keyframes to produce a temporally coherent video. The goal is to ensure smooth transitions between given keyframes.

Four keyframes are replicated to fill the first 7 latent frames, and the remaining 6 are interpolated.

- Clean condition: 4 frames (replicated keyframes)
- Buffer: 3 frames (noised condition)
- Target: 6 frames (pure noise)

We visualize the initial latent frames and their denoising process in Figure 11.

**Multiple Image Conditions** This task takes two distinct types of image conditions—such as a person and clothing, or a person and an object—and generates a target video that reflects the combination of

453 both. This setup is useful for applications like virtual try-on (VITON) or ad video synthesis, where  
454 two semantic entities must be jointly represented in motion.

455 The first 3 latent frames are derived from the first condition image, and the next 4 from the second  
456 condition image.

- 457 • Clean condition: 4 frames (3 from the first image, 1 from the second)
- 458 • Buffer: 3 frames (noised condition)
- 459 • Target: 6 frames (pure noise)

460 *Note that the number of condition sources is not limited to two; the framework supports arbitrary*  
461 *multi-condition setups.* We visualize the initial latent frames and their denoising process in Figure [12](#)

## 462 **A.6 Broader Impacts and Misuse Discussion**

463 Our TIC-FT method enables efficient adaptation of video diffusion models with minimal data.  
464 However, this ease of fine-tuning also introduces risks, particularly the potential misuse for creating  
465 deepfakes or misleading synthetic media. Clear usage policies and responsible deployment practices  
466 are essential to mitigate societal risks.

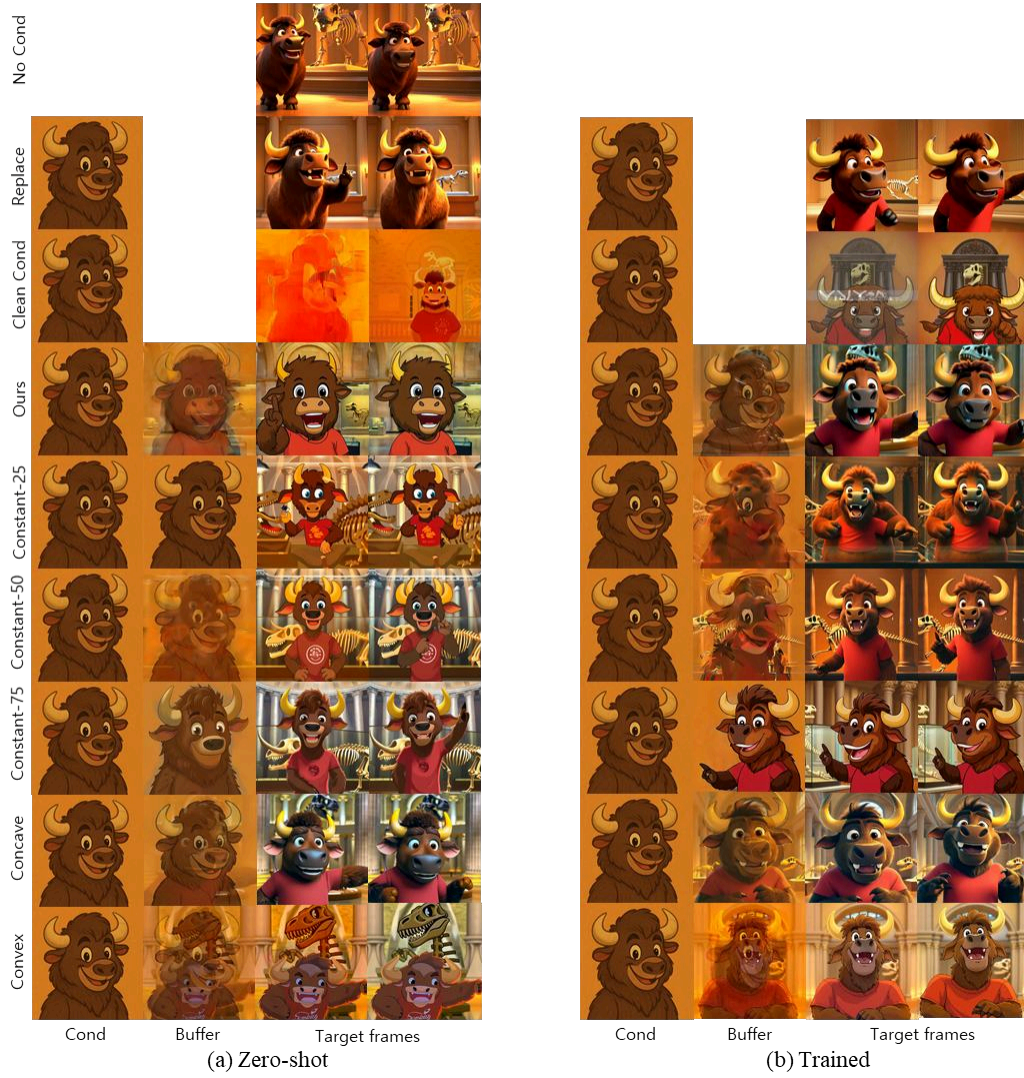


Figure 6: Qualitative comparison of buffer frame designs in zero-shot and fine-tuned settings.



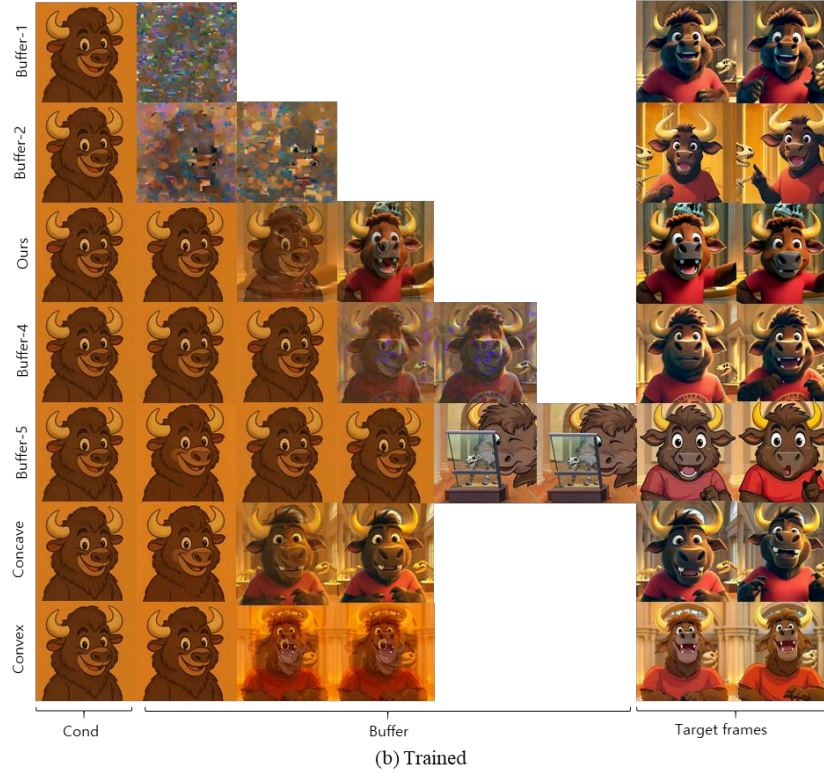
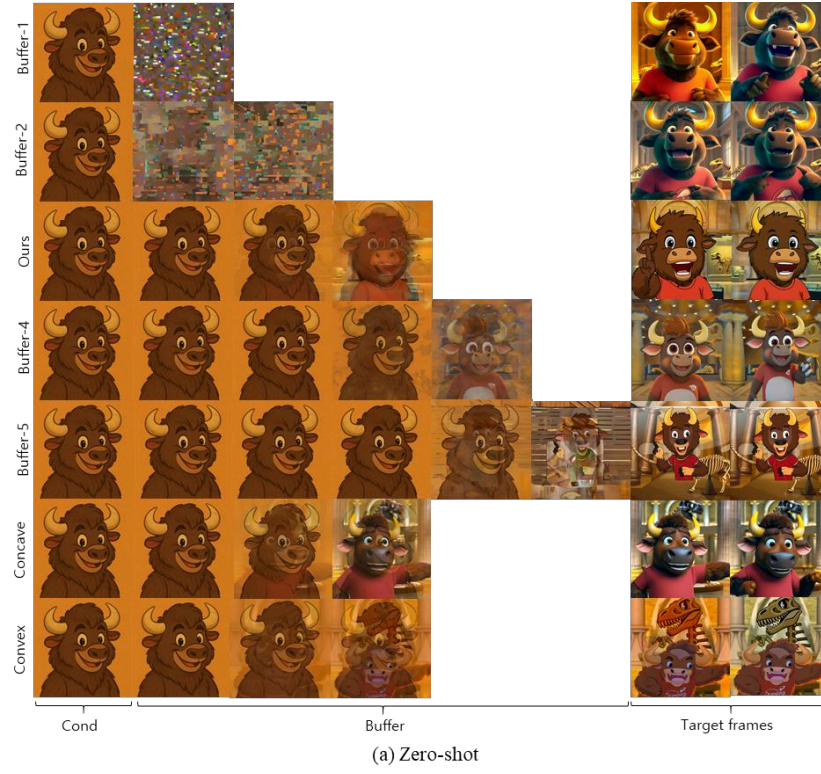


Figure 7: Qualitative comparison of buffer frame designs in zero-shot and fine-tuned settings.

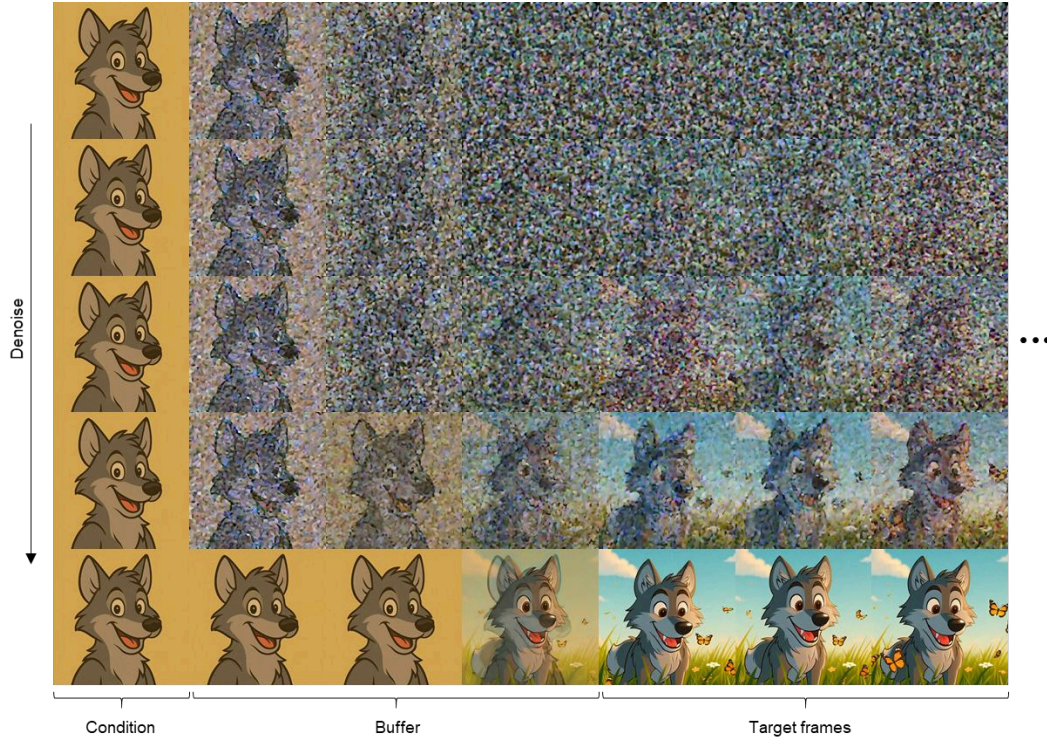


Figure 8: Visual results for initial frames and their denoising process on image-to-video generation.  
**Prompt:** *[Character] A clear, high-resolution front-facing close-up of a cheerful cartoon-style wolf character, centered against ...*

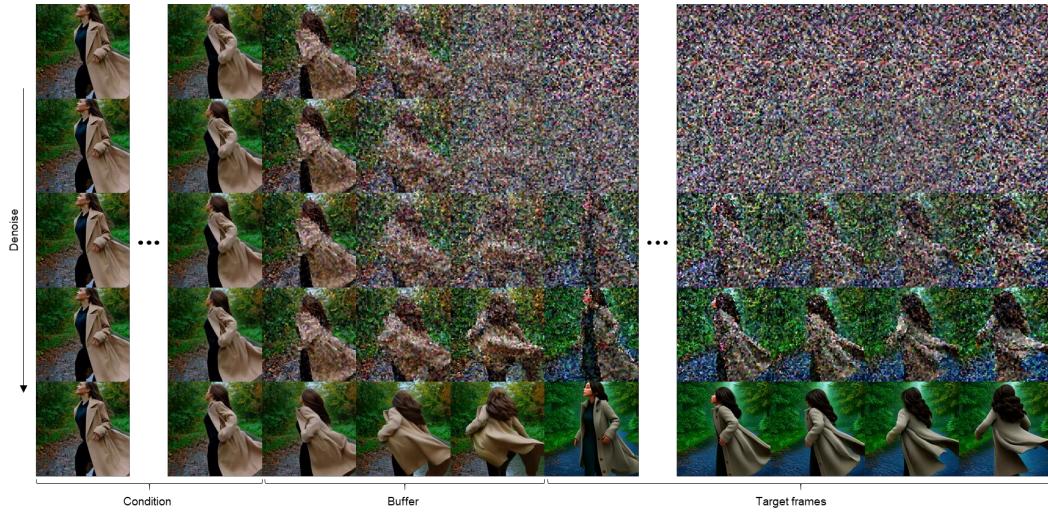


Figure 9: Visual results for initial frames and their denoising process on video style transfer task.  
**Prompt:** *[VIDEO1] A woman in a tan cloak walks gracefully along a forest path. Her hair flows gently with her movement, and the ...*



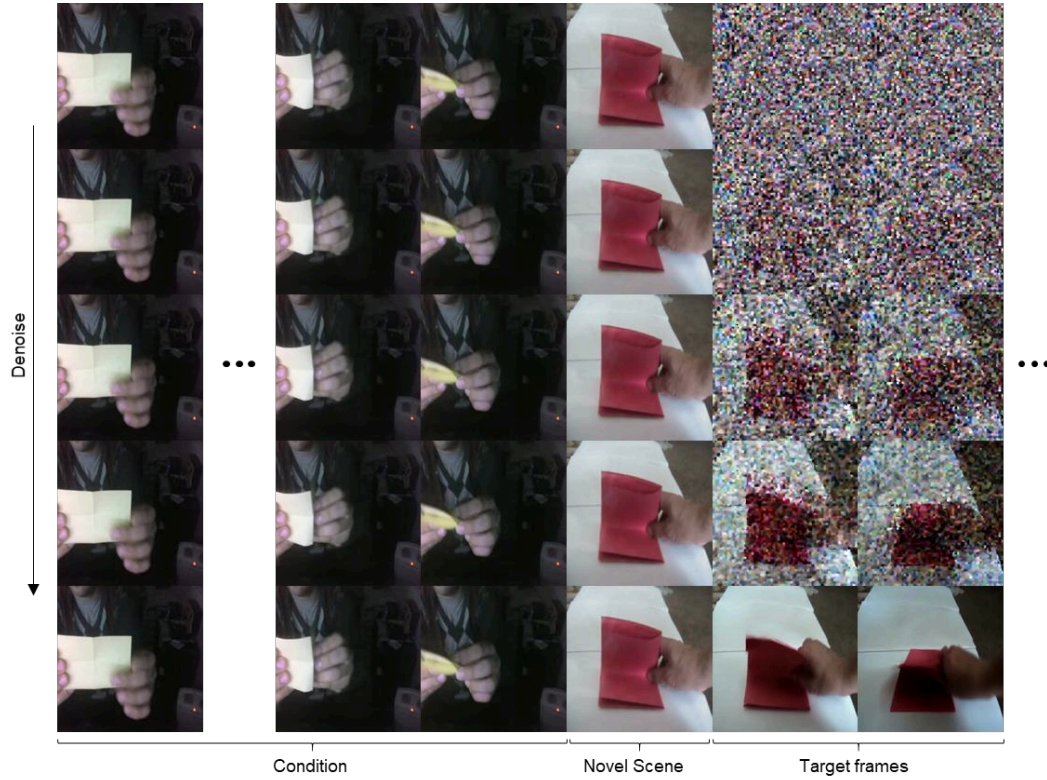


Figure 10: Visual results for initial frames and their denoising process on in-context action transfer task. **Prompt:** [REFERENCE VIDEO] A white paper is folded in half by a person wearing black sleeves in a dark indoor environment. ...



Figure 11: Visual results for initial frames and their denoising process on keyframe interpolation task. **Prompt:** [VIDEO1] A cartoon woman with red hair and a jeweled headpiece slowly tilts her head and changes facial expressions ...



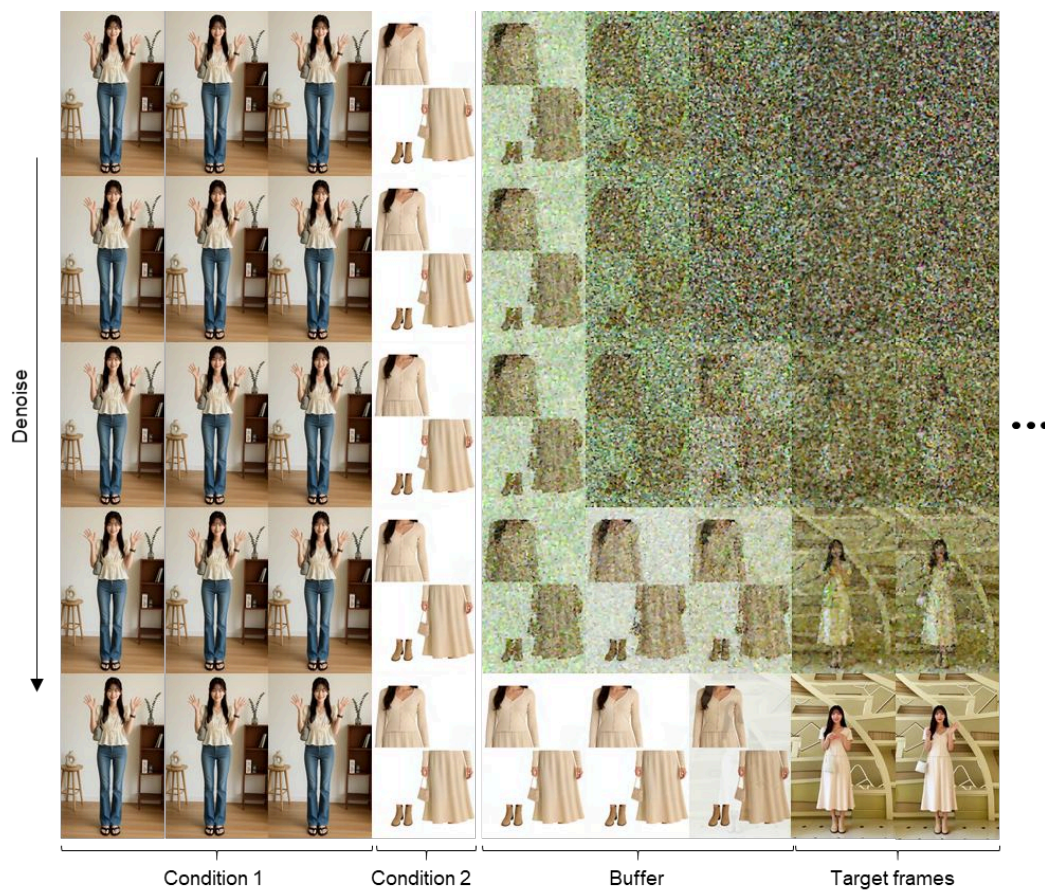


Figure 12: Visual results for initial frames and their denoising process on virtual try-on task. **Prompt:** *[IMAGE] A young woman with long black hair, wearing a cream blouse, blue jeans, and black sandals, smiles with both ...*