# VORCH-IR: LONG-FORM MULTIMODAL IDENTITY REPLACEMENT VIDEO GENERATION VIA A UNIFIED FRAMEWORK

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Identity replacement aims to replace subject identities in videos while preserving motion, expressions, and temporal consistency. Existing methods primarily focus on single-person scenarios and rely on task-specific structural guidance, such as masks, pose maps, or motion control signals, making them difficult to integrate into unified image-text-video editing frameworks. Moreover, the lack of large-scale paired training data limits progress on multi-person replacement. We propose a unified identity replacement framework that supports both single-person and dual-person replacement, with optional background replacement, within a single model. Built upon the LTX2 backbone, our method jointly conditions on textual instructions, reference identity images, optional background images, and driving videos. Notably, our approach eliminates the need for pose or spatial alignment in the reference inputs; multiple subject identities and background references can be fully decoupled and directly fed into the model. To support training, we develop an automatic data construction pipeline that synthesizes high-quality paired data for all editing settings. We further extend the model to long-form video generation and introduce Long-Horizon Error Control, enabling minute-scale video generation. Experiments demonstrate strong identity preservation, motion consistency, and temporal coherence across diverse replacement scenarios.

## 1 INTRODUCTION

Recent advances in diffusion-based video generation have substantially improved identity-preserving animation and replacement. Existing methods commonly formulate the task as pose-guided character animation, relying on masks, pose maps, or other structural signals to transfer a reference identity while preserving motion (Wang et al., 2025; Zhang et al., 2025; Shi et al., 2025; Yan et al., 2026b;a; Cheng et al., 2025). General video editing models such as VACE (Jiang et al., 2025) also support identity replacement as part of broader editing capabilities. However, existing approaches typically require task-specific structural guidance and focus primarily on single-person scenarios, while unified multi-person replacement remains underexplored, partly due to limited large-scale paired training data.

To address these limitations, we propose Vorch-IR, a unified identity replacement framework that supports four representative settings: single- and dual-person replacement with optional background editing. Built on the LTX2 (HaCohen et al., 2026) backbone, Vorch-IR jointly conditions on driving videos, identity references, optional background references, and textual instructions. Our decoupled input design removes pose, spatial, and layout constraints on reference images, enabling multiple identity and scene references to be incorporated within a unified architecture.

To enable large-scale training, we develop an automatic data construction pipeline that synthesizes paired replacement videos across all four settings, providing consistent motion and identity supervision without task-specific data collection. The proposed pipeline provides an efficient and scalable solution for generating diverse training data with consistent motion and identity supervision, making large-scale training of unified identity replacement models feasible.

We further extend Vorch-IR to long-form video generation. Since the model is trained on short clips, our approach adopts autoregressive continuation to generate long videos, which can accumulate



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

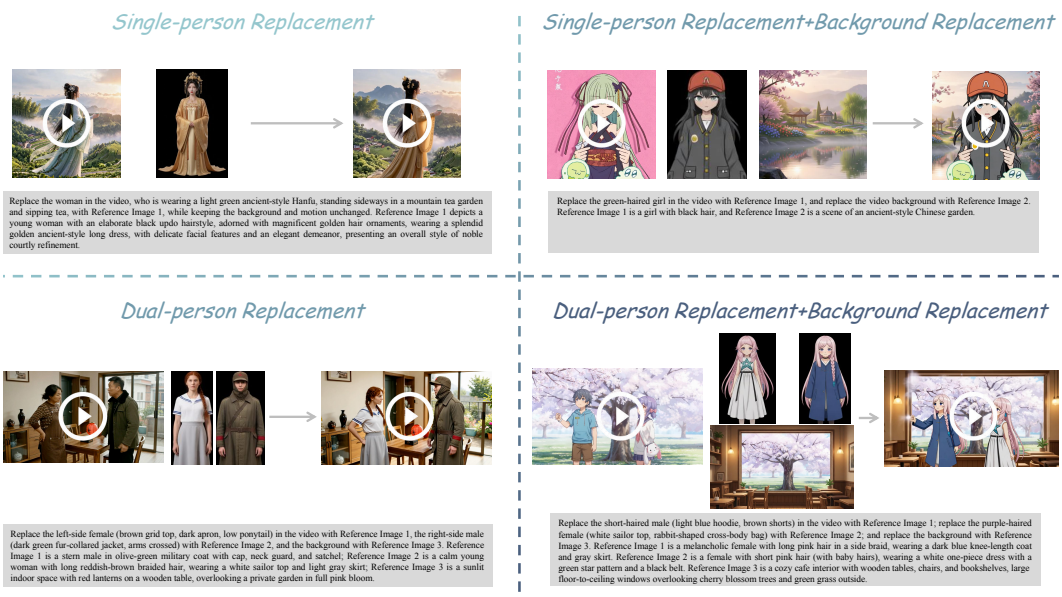



Figure 1: **Overview of four identity replacement settings.** A single model performs single-person and dual-person replacement, with optional background replacement. The textual instruction associates each indexed reference image with a subject or the background in the driving video.

errors over long temporal horizons. We therefore introduce Long-Horizon Error Control (LHEC), which combines Error-Aware learning (EA) during training with Temporal Test-time Correction (TTC) during inference to mitigate error accumulation and maintain stable identity replacement over minute-long videos.

Experimental results demonstrate that Vorch-IR achieves high-quality identity replacement while preserving motion dynamics and scene consistency across diverse editing scenarios. Our model generalizes effectively to both single-person and dual-person replacement, demonstrating strong performance across different replacement tasks.

Our contributions are summarized as follows:

- We formulate identity replacement as a unified multimodal conditional video generation problem, supporting single- and dual-person replacement with optional background editing through text-guided instructions.
- We introduce a completely decoupled input mechanism that eliminates the need for pose or spatial alignment in reference images, enabling multiple independent identity and scene references within a unified architecture.
- We develop a scalable data construction pipeline for automatically generating paired training data across diverse replacement settings, addressing the critical bottleneck of data scarcity.
- We introduce Long-Horizon Error Control (LHEC), combining Error-Aware learning (EA) and Temporal Test-time Correction (TTC) to enable stable identity replacement over minute-long videos.

## 2 Related Work

### 2.1 Video Generation

Diffusion models have become a dominant paradigm for video generation, with systems such as LaVie and SEINE demonstrating high-quality latent-space synthesis and flexible temporal generation (Wang et al., 2024; Chen et al., 2024b). Latte introduced a latent Diffusion Transformer for

text-to-video generation, establishing a scalable transformer design for subsequent video foundation models (Ma et al., 2025a). CogVideoX further develops this direction with an expert transformer tailored to text-conditioned synthesis (Yang et al., 2024). More recent large-scale systems, including HunyuanVideo, Wan, and Seedance, improve motion modeling, prompt following, and visual fidelity through advances in architecture, data curation, scaling, and post-training (Tencent Hunyuan Team, 2024; Wan Team, 2025; Gao et al., 2025). These models provide strong generative priors, but are not specifically designed to bind multiple spatially unaligned identity and scene references to targets in a driving video.

### 2.2 Identity Replacement and Animation

Reference-conditioned video diffusion models have substantially advanced identity-preserving animation and replacement (Ma et al., 2025b; 2026). Existing methods commonly rely on structural or spatial guidance, including masks, pose signals, and reference-conditioned features (Hu et al., 2025b; Cheng et al., 2025; Xu et al., 2026; Yan et al., 2026a). Recent works further explore environment-aware animation and multi-character generation (Hu et al., 2025a; Luo et al., 2026; Hu et al., 2026). In contrast, Vorch-IR uses a unified indexed interface for identity and background references and natural-language instructions to specify their target roles, without requiring spatially aligned references or subject masks.

### 2.3 Diffusion-Based Video Editing

General-purpose diffusion models provide another route to identity replacement. VACE supports a broad range of video creation and editing tasks by combining reference images with spatial controls that localize the edited regions (Jiang et al., 2025). AnyV2V offers a tuning-free alternative that edits the first frame and propagates the resulting appearance through an image-to-video diffusion model (Ku et al., 2024). Although flexible, these pipelines rely on spatial controls or first-frame propagation and do not explicitly address instruction-grounded correspondence among multiple identity and background references.

## 3 Method

We formulate the four replacement settings as a single multimodal conditional generation problem. We first describe the shared conditioning interface and its implementation in the LTX2 backbone (Sec. 3.1), then present the automatic data construction pipeline (Sec. 3.2). Finally, Sec. 3.3 introduces long-form video generation for minute-long sequences.

### 3.1 Unified Identity Replacement Framework

Given a driving video $D$, an ordered set of reference images $\mathcal{R} = \{R_1, \dots, R_M\}$, and an editing instruction $u$, Vorch-IR generates a target video that follows the motion of $D$ while transferring the specified identities and, optionally, the background. The same interface supports all four settings in Fig. 1, where $u$ specifies the semantic role of each indexed reference without requiring spatial alignment or task-specific networks.

**Instruction-Grounded Reference Binding.** Let $D$ denote the driving video, let $\mathcal{R} = \{R_1, \dots, R_M\}$ denote all available reference images, and let $u$ denote the editing instruction. The set $\mathcal{R}$ uniformly includes identity references and, when present, a background reference. Each image is identified only by its ordered reference index; whether $R_j$ represents a subject or the background is specified by $u$.

To jointly understand the driving content, reference appearances, and editing relations, we construct a multimodal prompt from the instruction, up to five representative frames $\mathcal{S}(D) = \{D_1, \dots, D_K\}$ sampled from the driving video, and all indexed references:

$$P = \mathrm{Template}\big(u, \{\langle \text{frame } k\rangle D_k\}_{k=1}^{K}, \{\langle \text{ref } j\rangle R_j\}_{j=1}^{M}\big)\,. \tag{1}$$

Here, the textual markers denote the ordering of image placeholders in the task template; no special background marker is required. A vision-language model $\mathcal{G}$ jointly contextualizes the text and

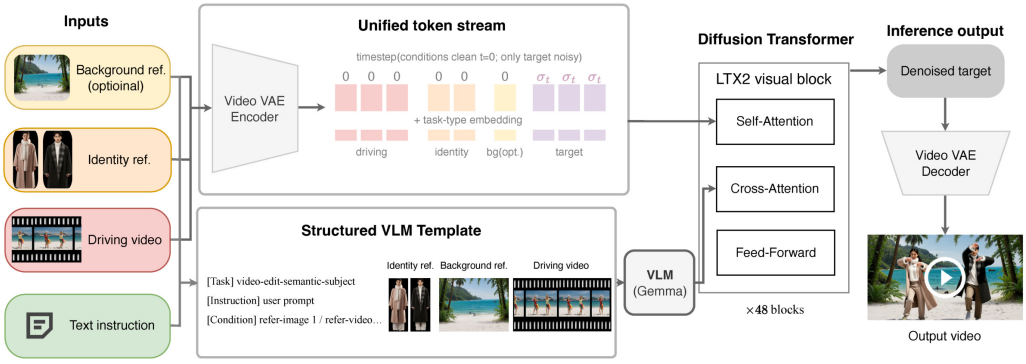



Figure 2: **Overview of Vorch-IR.** The driving video, indexed reference images, and noisy target latent form a unified visual token stream with task-ID embeddings. A Gemma vision-language encoder jointly processes the instruction, sampled driving frames, and references to produce cross-attention context. A shared Diffusion Transformer combines frame-aligned motion guidance with instruction-grounded reference appearance.

images, and an embedding connector $\mathcal{C}$ maps its hidden states to the context space of the diffusion transformer:

$$H = \mathcal{C}(\mathcal{G}(P)) \in \mathbb{R}^{L_c \times d_c}. \tag{2}$$

The resulting context $H$ jointly represents the source content, reference semantics, and instruction-defined roles. Reference-to-target correspondence is therefore learned as a semantic relation rather than produced by an explicit localization or matching module.

**Unified Visual-Semantic Attention.** Let $x_t$ denote the noisy target latent at diffusion timestep $t$, $\mathcal{E}$ the shared video VAE, and $\mathcal{P}$ latent patchification followed by the transformer input projection. We construct a unified visual token stream by concatenating the driving video, all reference conditions, and the noisy target:

$$S \equiv Z^0 = \left[ \mathcal{P}(\mathcal{E}(D)) \parallel \Big\|_{j=1}^{M} \mathcal{P}(\mathcal{E}(R_j)) \parallel \mathcal{P}(x_t) \right], \tag{3}$$

where $\parallel$ denotes token-wise concatenation. At each transformer block, joint self-attention is performed over the entire visual stream. Consequently, every target token can directly aggregate fine-grained appearance, spatial structure, and motion information from the driving video and all reference inputs. This dense interaction enables frame-aligned information transfer from the driving video while preserving detailed visual evidence from multiple identity and background references.

The self-attention output is subsequently conditioned on the multimodal VLM context $H$ in Eq. equation 2 through cross-attention. The context jointly encodes the editing instruction, sampled frames from the driving video, and the ordered reference images. It therefore facilitates learned semantic routing between each indexed reference and its instruction-specified subject or background role. In this design, self-attention provides dense visual content fusion, whereas VLM cross-attention supplies instruction-grounded semantic binding. Only the target-token slice is retained for the denoising prediction, while the driving and reference tokens serve as visual conditions. Adding or removing an identity or background condition only changes the corresponding reference-token groups, allowing the same architecture and shared parameters to support different combinations of visual conditions.

**Spatiotemporal and Role Encoding for Multi-Reference Visual Conditioning** To distinguish different conditioning roles, we introduce learnable task embeddings $E_{\text{task}} \in \mathbb{R}^{K \times D}$, where $K$ denotes the number of supported task IDs and $D$ is the transformer hidden dimension. For each visual token $\mathbf{z}_i$, we add the embedding $\mathbf{e}_{c_i}$ corresponding to its task ID $c_i$: $\mathbf{x}_i = W_{\text{patch}}\mathbf{z}_i + \mathbf{e}_{c_i}$. The resulting task-aware representations are propagated through all subsequent transformer layers, providing an additional condition-identity signal that distinguishes different references even when

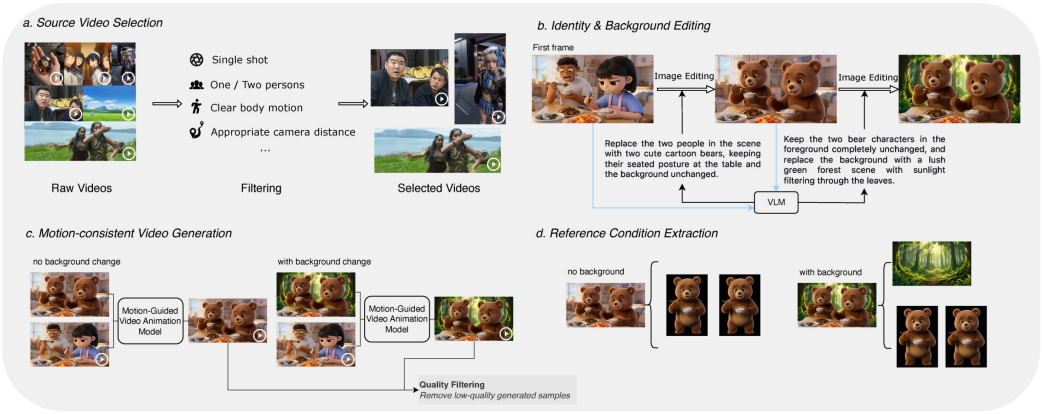



Figure 3: **Automatic data construction.** We edit the identities and, when required, the background of the first frame; propagate the edited appearance through the source motion; filter failed generations; and extract identity and background references to form paired training tuples.

they share the same RoPE coordinate system. Details of the task embeddings are provided in the supplementary material.

For long-video generation, the continuation context and denoising region share the same factorized 3D RoPE with continuous temporal coordinates:

$$\rho_{\text{cont}} = ([0, T_c), [0, H), [0, W)) \qquad \rho_{\text{denoise}} = ([T_c, T_v), [0, H), [0, W)) \,, \tag{4}$$

where $T_c$ is the temporal length of the continuation context and $T_v$ is the total length of the current video window. As shown in Eq. equation 4, the continuation and denoising regions use continuous temporal coordinates. The continuation context occupies $[0, T_c)$ and provides clean latent states as fixed context, while the denoising region occupies $[T_c, T_v)$ and is represented by the current noisy state. During denoising, a binary mask restricts the diffusion update to the denoising region, leaving the continuation context unchanged. This design allows RoPE to encode their spatiotemporal continuity while the diffusion mask separates the fixed context from the generated region.

### 3.2 Data Construction Pipeline

Large-scale paired videos that preserve source motion while modifying one or more identities and, optionally, the background are difficult to obtain manually. We therefore construct training pairs automatically from raw human videos. As illustrated in Fig. 3, each training sample consists of a driving video, one or more identity references, an optional background reference, an editing instruction, and the corresponding target video.

We collect human videos from public datasets and internal collections, and retain single-shot clips containing one or two clearly visible subjects with recognizable faces and meaningful body motion. For each selected clip, an off-the-shelf image editor Qwen-Image-Edit Wu et al. (2025) modifies the identities in the first frame while approximately preserving their locations and coarse poses. The synthesized references cover realistic people, anime and cartoon characters, and other humanoid appearances. For background-editing samples, the scene is additionally replaced with a semantically compatible background. A large language model then generates the corresponding editing instruction, specifying the role of each indexed reference.

The edited first frame and original clip are subsequently fed into a motion-guided video animation model SCAIL2 Yan et al. (2026a), which propagates the edited identities and background along the original motion trajectories. A vision-language model filters the generated results for common failures, including malformed bodies, incorrect subject counts, identity mismatch, and unsuccessful background replacement. Only samples that preserve the intended content and motion are retained.

Finally, we extract the edited subjects from the synthesized first frame as identity references. For background-editing samples, foreground removal is used to obtain a clean background reference.

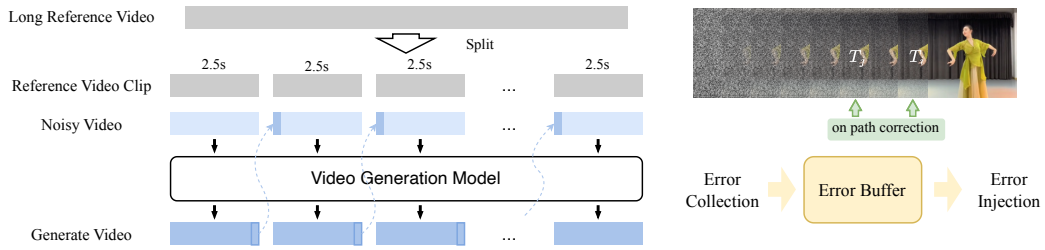



Figure 4: **Overview of the proposed Long-Form Video Generation Framework.** The figure illustrates the sequential generation architecture, the proposed Temporal Test-time Correction (TTC) for on-path error correction, and the Error-Aware learning strategy (EA) for improving robustness to accumulated errors during long-horizon generation.

Together with the driving video, editing instruction, and generated target, these components form the final training tuple. By independently varying the number of subjects and whether background editing is applied, the same pipeline covers all four replacement settings without task-specific collection or annotation procedures.

### 3.3 Long-Sequence Genreation

Although Vorch-IR is trained on short clips, practical identity replacement may require minute-long outputs. A common solution is autoregressive continuation, where each clip is conditioned on frames generated by the previous clip. Such feedback can accumulate errors and lead to identity drift or temporal degradation over long horizons (Chen et al., 2024a; Zhou et al., 2025; Huang et al., 2025).

To address this issue, we propose a long-form video generation framework shown in Fig. 4. To improve generation stability over long horizons, we introduce a Long-Horizon Error Control (LHEC) strategy that combines complementary training-based and training-free techniques to mitigate error accumulation and propagation. This enables longer-duration identity replacement while maintaining consistent character appearance, thereby extending the applicability of our approach to a broader range of long-form video editing scenarios.

**Long-Form Video Generation Architecture.** Given a long driving video, we first divide it into temporal chunks ${V^{(k)}}_{k=1}^{K}$ of equal length. Each chunk is processed sequentially through the full diffusion sampling process. The first chunk is generated normally using its corresponding driving-video segment and the global reference conditions. After decoding the generated result, we retain its last 25 frames as temporal context for the next chunk. Specifically, for $k > 1$, the current chunk is conditioned on the preceding generated frames:

$$\tilde{V}^{(k)} = \mathrm{Concat}_{\mathrm{time}}\left(\mathrm{Tail}_{25}\left(\hat{V}^{(k-1)}\right), V^{(k)}\right),$$

where $\hat{V}^{(k-1)}$ denotes the generated video chunk from the previous inference step. The context frames are concatenated with the current chunk and share the same temporal positional encoding scheme, preserving the temporal correspondence across consecutive chunks. The current chunk is then generated using $\tilde{V}^{(k)}$, the corresponding driving-video segment, and the same global reference conditions, after which its last 25 frames are retained as context for the next chunk.

**Long-Horizon Error Control.** To enable stable video generation over extended horizons, we introduce Long-Horizon Error Control (LHEC), which consists of two complementary components: Error-Aware learning (EA) and Temporal Test-time Correction (TTC). EA is inspired by Matrix-Game 3.0 Wang et al. (2026), where we sample normally generated video segments during training and compute their deviation $E$ from the corresponding ground-truth segments. The resulting error signals are stored in an error buffer and added to the reference frames during continuation training,

explicitly exposing the model to error-corrupted contexts and enabling it to learn to mitigate error accumulation.

TTC is inspired by Pathwise Test-Time Correction Xiang et al. (2026), which performs on-path correction during inference to compensate for drift accumulated along the autoregressive generation trajectory. At selected flow timesteps $t$, we first obtain a clean prediction from the non-continuation model $G_{\mathrm{NC}}$ without the autoregressive context $C^{(k)}$ and re-noise it to the same timestep, after which the continuation model $G_{\mathrm{C}}$ resumes the denoising process:

$$\begin{aligned} \hat{U}_{0,\mathrm{NC}}^{(k)} &= G_{\mathrm{NC}}\left(U_t^{(k)}, V^{(k)}, R\right), \qquad \tilde{U}_t^{(k)} = (1-t)\hat{U}_{0,\mathrm{NC}}^{(k)} + t\epsilon_t^{(k)}, \\ \hat{U}_{0,\mathrm{C}}^{(k)} &= G_{\mathrm{C}}\left(\tilde{U}_t^{(k)}, C^{(k)}, V^{(k)}, R\right), \qquad \epsilon_t^{(k)} \sim \mathcal{N}(0, I). \end{aligned} \tag{5}$$

Here, $U_t^{(k)}$ denotes the current noisy latent, while $\hat{U}_{0,\mathrm{NC}}^{(k)}$ and $\hat{U}_{0,\mathrm{C}}^{(k)}$ denote the clean latent predictions from the non-continuation and continuation models, respectively. The non-continuation prediction provides a temporally anchored estimate that is reintroduced into the ongoing flow trajectory, allowing the continuation model to mitigate accumulated autoregressive drift.

# 4 Experiments

We evaluate Vorch-IR across four editing settings and compare with representative methods in the shared single-person setting. Because competing methods do not expose a common interface for dual-person and background replacement, the remaining settings are evaluated qualitatively through results from our unified model. Dual-person and long-video replacement are further evaluated on our own benchmarks using automatic metrics and GSB human evaluation.

## 4.1 Implementation Details

**Datasets.** We combine public human-motion datasets with large-scale internal video collections. Public data primarily contribute dance sequences with complex body motion, whereas the internal data cover a broader range of everyday activities and scenes. We process all source videos with the pipeline in Sec. 3.2 and divide long videos into clips of at most 8 seconds. Approximately 91% of the clips are between 2 and 4 seconds long. The resulting training set contains 15,571 clips totaling approximately 13 hours, with both single- and dual-person scenes under diverse motions and backgrounds.

**Training details.** We initialize the model from the official LTX2 checkpoint and fine-tune the Diffusion Transformer, including its task-ID embeddings, while keeping the Gemma vision-language encoder and embedding connector frozen. We use the checkpoint after 2,000 optimizer updates and train on 32 NVIDIA H20Z GPUs with a per-device batch size of 1 and three-step gradient accumulation, yielding an effective global batch size of 96. We use AdamW with a learning rate of $1 \times 10^{-5}$, sample noise levels from a shifted logit-normal distribution, and drop the textual condition with probability 0.1. During training, each sample has a 50% probability of being converted to the continuation setting, where the first 25 frames of the ground-truth video are encoded with the VAE and prepended to the noisy target tokens as continuation context, while denoising is performed only on the subsequent frames. We further apply the proposed Error-Aware learning (EA) strategy by augmenting continuation samples with error-corrupted contexts following Sec. 3.3.

**Long-Video Inference.** The long-sequence inference pipeline divides the driving video into 2.5-second chunks and processes them sequentially with a 20-step sampler. The first chunk is generated using its corresponding driving-video segment and global reference conditions. For each subsequent chunk, the previous generated chunk is decoded into pixel space, and its last 25 frames are re-encoded by the VAE as temporal context. This decode-and-re-encode step is required by the causal temporal structure of LTX-VAE, which prevents direct reuse of the previous latent chunk. TTC is applied at denoising steps 16 and 17, where the non-continuation prediction provides an additional temporal anchor to mitigate error accumulation across successive chunks. The model is trained only on short clips and does not require full-length training sequences.

| Methods | Motion smoothness ↑ | Imaging quality ↑ | Temporal flickering ↑ | Subject consistency ↑ | Background consistency ↑ | Aesthetic quality ↑ |
|---|---|---|---|---|---|---|
| Wan-Animate | 0.9828 | 0.6732 | 0.9774 | 0.8787 | 0.9184 | 0.5756 |
| HunyuanCustom | **0.9861** | 0.6809 | **0.9781** | 0.9002 | 0.9328 | 0.5745 |
| MoCha | 0.9792 | 0.6675 | 0.9682 | 0.8935 | 0.9102 | 0.6175 |
| SCAIL-2 | 0.9750 | **0.7303** | 0.9617 | 0.8622 | 0.9145 | **0.6602** |
| Vorch-IR (Single) | 0.9814 | 0.7031 | 0.9680 | **0.9096** | **0.9332** | 0.6396 |
| Vorch-IR (Dual) | 0.9792 | 0.7083 | 0.9629 | 0.8761 | 0.9084 | 0.6012 |

Table 1: **VBench results on our benchmark.** Existing methods are evaluated in the single-person setting, while Vorch-IR is additionally evaluated in the dual-person setting.

| Methods | Subject Similarity | | Motion Consistency |
|---|---|---|---|
| | DINOv2 ↑ | CLIP-I ↑ | DWPose error ↓ |
| Wan-Animate | 0.4411 | 0.7434 | **0.0614** |
| HunyuanCustom | 0.4379 | 0.7491 | 0.0901 |
| MoCha | 0.4634 | 0.7504 | 0.0957 |
| SCAIL-2 | 0.5470 | **0.7835** | 0.1012 |
| Vorch-IR (Single) | **0.5613** | 0.7767 | 0.0976 |
| Vorch-IR (Dual) | 0.3794 | 0.7143 | 0.0803 |

Table 2: **Subject similarity and motion consistency on our benchmark.** Single-person results for all methods and dual-person results for Vorch-IR.

| Methods | Motion smoothness ↑ | Imaging quality ↑ | Temporal flickering ↑ |
|---|---|---|---|
| Wan-Animate | 1.3245 | 3.9663 | **3.6635** |
| HunyuanCustom | 1.2716 | 3.1078 | 3.4052 |
| MoCha | 1.1786 | 4.2405 | 3.3452 |
| SCAIL-2 | 1.1639 | **4.8215** | 2.7855 |
| Vorch-IR | **1.4286** | 4.2786 | 3.3333 |

Table 3: **Video-Bench results on the XDance benchmark.** We follow the original benchmark protocol and evaluate all methods under the single-person replacement setting.

### 4.2 Evaluation

We compare Vorch-IR with Wan-Animate, HunyuanCustom, MoCha, and SCAIL-2 in the shared single-person replacement setting. Since existing methods are designed for single-reference inputs, we restrict comparisons to this setting without modifying their interfaces. Limitations and failure cases of Vorch-IR are discussed in the supplementary material. Our demos are also provided in the supplementary material as an interactive HTML webpage.

**General Video Generation.** We complement the visual comparison with quantitative results on our benchmark and XDance, covering both single- and dual-subject replacement tasks. Tab. 1 and 2 report automatic metrics on our benchmark, while Tab. 3 evaluates cross-dataset generalization on XDance. Higher values indicate better results for all metrics except DWPose error, for which lower values are better. More qualitative results are presented in the supplementary material.

As shown in Tab. 1, Vorch-IR achieves the best subject consistency 0.9096 and background consistency 0.9332, while maintaining competitive visual and temporal quality. It also obtains strong motion smoothness 0.9814, close to the best result 0.9861. Tab. 2 further shows that Vorch-IR achieves the highest DINOv2 similarity (0.5613) and second-best CLIP-I score (0.7767). On XDance (Tab. 3), it achieves the best motion smoothness (1.4286) and second-best imaging quality (4.2786). Overall, Vorch-IR demonstrates strong identity fidelity with competitive motion and visual quality across benchmarks.

Since existing comparisons are limited to the shared single-person setting, we further evaluate Vorch-IR's ability to perform dual-person replacement with optional background editing using a dedicated benchmark. The results are reported in the last rows of Tab. 1 and Tab. 2. Additional qualitative examples provided in the supplementary material. We further conduct a pairwise Good–Same–Bad (GSB) human study, comparing Vorch-IR with each competing method under identical inputs. As shown in Fig. 5, Vorch-IR consistently obtains a higher Good than Bad share across evaluated dimensions, with the largest advantage observed in identity consistency. Comparisons with SCAIL-2 yield a higher proportion of same judgments, particularly for physical plausibility and video quality. These results complement automatic metrics by capturing perceptual differences in complete videos.

**Long-Video Generation.** To analyze the effectiveness of EA and TTC for long-video generation, we conduct ablation studies on our long-video benchmark covering all defined sub-tasks. We

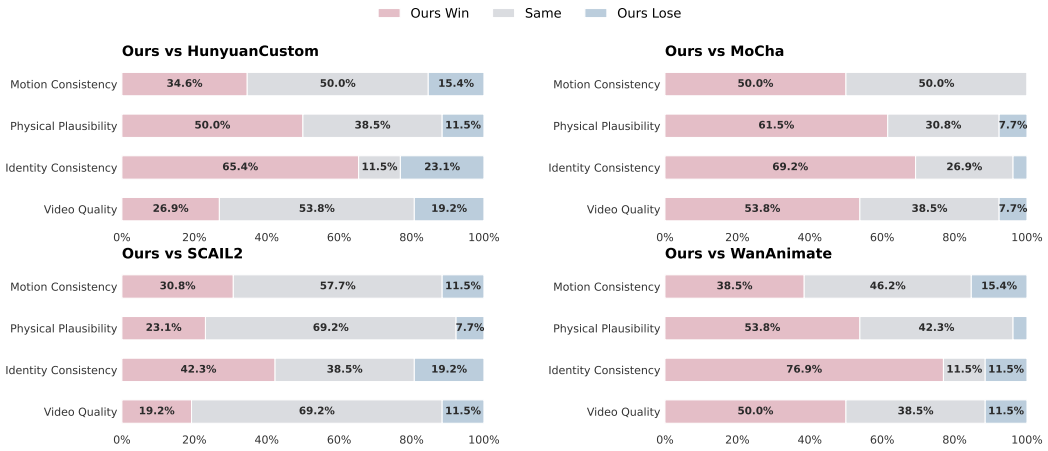



Figure 5: **Good–Same–Bad (GSB) pairwise human evaluation.** For each dimension and baseline, Good indicates that evaluators prefer Vorch-IR, Same indicates comparable perceptual quality, and Bad indicates a preference for the competing method. The numbers are percentages, and each distribution is normalized to 100%.

| Methods | Motion smoothness ↑ | Imaging quality ↑ | Temporal flickering ↑ | Subject consistency ↑ | Background consistency ↑ | Aesthetic quality ↑ | Reference identity ↑ | Flow residual ↓ |
|---|---|---|---|---|---|---|---|---|
| Ours (Full) | **0.9910** | **0.6874** | **0.9777** | **0.8344** | **0.8812** | **0.6270** | **0.3085** | **0.0112** |
| Ours w/o EA | 0.9898 | 0.6730 | 0.9749 | 0.8269 | 0.8716 | 0.6232 | 0.2633 | 0.0129 |
| Ours w/o TTC | 0.9900 | 0.6572 | 0.9747 | 0.8201 | 0.8744 | 0.6106 | 0.2923 | 0.0132 |
| Ours w/o EA & TTC | 0.9891 | 0.6472 | 0.9726 | 0.8208 | 0.8708 | 0.6158 | 0.2594 | 0.0143 |

Table 4: **Ablation study on the methods used for TTC on our long-video benchmark.** We compare the full model with variants without TTC (w/o TTC), without EA (w/o EA), and without both TTC and EA (w/o TTC & EA). Results demonstrate the contribution of each component to long-video generation.

use selected VBench Huang et al. (2024) metrics together with two task-specific metrics: Reference identity, measuring similarity between generated subjects and references, and Flow residual, measuring motion discrepancy across consecutive segments. Different TTC step configurations are further studied in the supplementary material, where the 16–17 setting is selected as the default configuration.

As shown in Tab. 4, the full model achieves the best overall performance among different configurations. Removing EA or TTC degrades performance, with the largest drop occurring when both are removed. In particular, Reference identity decreases from 0.3085 to 0.2594, while Flow residual increases from 0.0112 to 0.0143, demonstrating the effectiveness of EA and TTC in preserving identity and temporal consistency.

## 5 Conclusion

We presented Vorch-IR, a unified multimodal framework for single- and dual-person video identity replacement with optional background editing. The model represents the driving video, identity references, background reference, and noisy target in a shared visual token stream, while a vision-language context binds indexed references to instruction-defined targets. This design requires neither spatially aligned references nor masks or pose inputs at inference time, enabling flexible identity and scene replacement within a unified architecture. An automatic construction pipeline supplies paired training data for all four settings, and a shared flow-matching objective trains them within one model. Finally, we introduce Long-Horizon Error Control (LHEC), which combines Error-Aware learning (EA) and Temporal Test-time Correction (TTC) to extend short-clip generation models to minute-long videos through window-level prediction and full-sequence diffusion updates, effectively maintaining identity preservation, motion consistency, and temporal coherence over long horizons.

### AI USE STATEMENT

During the preparation of this work, we used Google Gemini to assist with dataset cleaning and quality-control filtering, WorkBuddy and OpenAI GPT to assist with literature searches and preliminary summaries of potentially relevant papers, and OpenAI GPT to improve the clarity and readability of the paper. All AI-assisted outputs were reviewed and verified by the authors. The final methodology, experimental design, analysis, and conclusions were reviewed and verified by the authors. The authors take full responsibility for the final content of this work, including all text, claims, data, analyses, and artifacts produced with the assistance of generative AI.

### REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the model architecture, training procedure, data construction, and evaluation protocols in the main paper and supplementary material. We also provide training configurations, implementation details, ablation studies, and additional qualitative and quantitative results to support the reproduction, evaluation, and further study of our method.

## REFERENCES

Boyuan Chen, Diego Marti Monso, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *arXiv preprint arXiv:2407.01392*, 2024a.

Xinyuan Chen, Yaohui Wang, Lingjun Zhang, Shaobin Zhuang, Xin Ma, Jiashuo Yu, Yali Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Seine: Short-to-long video diffusion model for generative transition and prediction. In *International Conference on Learning Representations*, 2024b.

Gang Cheng, Xin Gao, Li Hu, Siqi Hu, Mingyang Huang, Chaonan Ji, Ju Li, Dechao Meng, Jinwei Qi, Penchong Qiao, et al. Wan-animate: Unified character animation and replacement with holistic replication. *arXiv preprint arXiv:2509.14055*, 2025.

Yu Gao, Haoyuan Guo, Tuyen Hoang, Weilin Huang, Lu Jiang, Fangyuan Kong, Huixia Li, Jiashi Li, Liang Li, Xiaojie Li, Xunsong Li, Yifu Li, Shanchuan Lin, Zhijie Lin, Jiawei Liu, Shu Liu, et al. Seedance 1.0: Exploring the boundaries of video generation models. *arXiv preprint arXiv:2506.09113*, 2025.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, et al. Ltx-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2026.

Li Hu, Guangyuan Wang, Zhen Shen, Xin Gao, Dechao Meng, Lian Zhuo, Peng Zhang, Bang Zhang, and Liefeng Bo. Animate anyone 2: High-fidelity character image animation with environment affordance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10207–10217, 2025a.

Teng Hu, Zhentao Yu, Zhengguang Zhou, Sen Liang, Yuan Zhou, Qin Lin, and Qinglin Lu. Hunyuancustom: A multimodal-driven architecture for customized video generation. *arXiv preprint arXiv:2505.04512*, 2025b.

Yingcheng Hu, Haowen Gong, Chuanguang Yang, Zhulin An, Yongjun Xu, and Songhua Liu. Multianimate: Pose-guided image animation made extensible. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9306–9316, 2026.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*, 2025.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818. IEEE, 2024.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17191–17202, 2025.

Max Ku, Cong Wei, Weiming Ren, Harry Yang, and Wenhu Chen. Anyv2v: A tuning-free framework for any video-to-video editing tasks. *arXiv preprint arXiv:2403.14468*, 2024.

Mingshuang Luo, Shuang Liang, Zhengkun Rong, Yuxuan Luo, Tianshu Hu, Ruibing Hou, Hong Chang, Yong Li, Yuan Zhang, and Mingyuan Gao. Dreamactor-m2: Universal character image animation via spatiotemporal in-context learning. *arXiv preprint arXiv:2601.21716*, 2026.

Xin Ma, Yaohui Wang, Xinyuan Chen, Gengyun Jia, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *Transactions on Machine Learning Research*, 2025a.

Xin Ma, Yaohui Wang, Gengyun Jia, Xinyuan Chen, Tien-Tsin Wong, Yuan-Fang Li, and Cunjian Chen. Consistent and controllable image animation with motion diffusion models. In *Computer Vision and Pattern Recognition*, pp. 7288–7298, 2025b.

Xin Ma, Yaohui Wang, Genyun Jia, Xinyuan Chen, Tien-Tsin Wong, and Cunjian Chen. Consistent and controllable image animation with motion linear diffusion transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Shijun Shi, Jing Xu, Zhihang Li, Chunli Peng, Xiaoda Yang, Lijing Lu, Kai Hu, and Jiangning Zhang. One-to-all animation: Alignment-free character animation and image pose transfer. *arXiv preprint arXiv:2511.22940*, 2025.

Tencent Hunyuan Team. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Wan Team. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Xiang Wang, Shiwei Zhang, Longxiang Tang, Yingya Zhang, Changxin Gao, Yuehuan Wang, and Nong Sang. Unianimate-dit: Human image animation with large-scale video diffusion transformer. *arXiv preprint arXiv:2504.11289*, 2025.

Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *International Journal of Computer Vision*, pp. 1–20, 2024.

Zile Wang, Zexiang Liu, Jiaxing Li, Kaichen Huang, Baixin Xu, Fei Kang, Mengyin An, Peiyu Wang, Biao Jiang, Yichen Wei, et al. Matrix-game 3.0: Real-time and streaming interactive world model with long-horizon memory. *arXiv preprint arXiv:2604.08995*, 2026.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025. URL `https://arxiv.org/abs/2508.02324`.

Xunzhi Xiang, Zixuan Duan, Guiyu Zhang, Haiyu Zhang, Zhe Gao, Junta Wu, Shaofeng Zhang, Tengfei Wang, Qi Fan, and Chunchao Guo. Pathwise test-time correction for autoregressive long video generation. *arXiv preprint arXiv:2602.05871*, 2026.

Zhengbo Xu, Jie Ma, Ziheng Wang, Zhan Peng, Jun Liang, and Jing Li. Mocha: End-to-end video character replacement without structural guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16279–16288, 2026.

Wenhao Yan, Fengjia Guo, Zhuoyi Yang, and Jie Tang. Scail-2: Unifying controlled character animation with end-to-end in-context conditioning. *arXiv preprint arXiv:2606.10804*, 2026a.

Wenhao Yan, Sheng Ye, Zhuoyi Yang, Jiayan Teng, ZhenHui Dong, Kairui Wen, Xiaotao Gu, Yong-Jin Liu, and Jie Tang. Scail: Towards studio-grade character animation via in-context learning of 3d-consistent pose representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Findings*, 2026b.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

Jiaming Zhang, Shengming Cao, Rui Li, Xiaotong Zhao, Yutao Cui, Xinglin Hou, Gangshan Wu, Haolan Chen, Yu Xu, Limin Wang, and Kai Ma. Steadydancer: Harmonized and coherent human image animation with first-frame preservation. *arXiv preprint arXiv:2511.19320*, 2025.

Deyu Zhou, Quan Sun, Yuang Peng, Kun Yan, Runpei Dong, Duomin Wang, Zheng Ge, Nan Duan, Xiangyu Zhang, Lionel M. Ni, and Heung-Yeung Shum. Taming teacher forcing for masked autoregressive video generation. *arXiv preprint arXiv:2501.12389*, 2025.

# A Appendix

## A.1 Additional Details of the Model Architecture

$$S \equiv Z^0 = \big[\, \mathcal{P}(\mathcal{E}(D)) + e_{\text{drive}} \,\|\, \Big\|_{j=1}^{M} \big(\mathcal{P}(\mathcal{E}(R_j)) + e_j\big) \,\|\, \mathcal{P}(x_t) + e_{\text{tgt}} \,\big], \tag{6}$$

In the unified token stream $S$, the task-ID embedding distinguishes the driving video and ordered reference slots: in our implementation, the driving video segment uses task ID 65, the $j$-th reference uses task ID $j$, and the target uses task ID 0, whose embedding is explicitly set to zero. These IDs identify input slots, not semantic identity/background categories; the role of $R_j$ is specified by $u$ and interpreted through the VLM context.

In our setting, the instruction $u$ may state that "Reference 1 replaces the person on the left, Reference 2 replaces the person on the right, and Reference 3 replaces the background." Thus, the ordered index identifies a reference slot, whereas the instruction defines its semantic role and target; no geometric correspondence is supplied.

## A.2 Data Construction

Our data construction pipeline incorporates several important considerations. First, the initial videos are selected according to the following criteria: (1) each video contains a single continuous shot; (2) the first frame contains one or two visible subjects; (3) the subjects are shown in medium shots, avoiding excessively close or distant views to preserve both clear body motion and sufficient appearance details; and (4) the subjects exhibit clear and meaningful body movements.

Second, after image editing, we ensure that the spatial positions of the subjects remain largely unchanged from those in the original frames. This constraint preserves the spatial consistency between the edited reference images and the original subject locations, which is important for the subsequent Motion-Guided Video Animation model to produce high-quality results.

## A.3 Evaluation Metrics and Results

Additional qualitative results are presented in Fig. 6 and Fig. 7. Tab. 5 reports an ablation study on the choice of sampling steps for TTC.

## A.4 Limitations and Failure Cases

Despite its unified formulation, Vorch-IR still exhibits several failure modes. First, reference grounding may occasionally fail, leading to incorrect identity assignment or unintended replacement of




Figure 6: **Single-person identity replacement comparison.** Qualitative results comparing our method with SCAIL2 and Wan Animate.



Figure 7: **Dual-person generation cases.** Qualitative results demonstrating the control of two identities and optional background replacement using the same model and input interface.

| TTC Steps | Motion smoothness ↑ | Imaging quality ↑ | Temporal flickering ↑ | Subject consistency ↑ | Background consistency ↑ | Aesthetic quality ↑ | Reference identity ↑ | Flow residual ↓ |
|---|---|---|---|---|---|---|---|---|
| 16–19 | 0.9909 | 0.6825 | 0.9777 | 0.8299 | 0.8784 | **0.6307** | 0.3056 | **0.0111** |
| 16–17 | 0.9910 | **0.6874** | 0.9777 | **0.8344** | **0.8812** | 0.6270 | **0.3085** | 0.0112 |
| 17–18 | **0.9911** | 0.6784 | **0.9781** | 0.8324 | 0.8757 | 0.6172 | 0.3013 | 0.0112 |
| 18–19 | 0.9906 | 0.6641 | 0.9762 | 0.8279 | 0.8787 | 0.6117 | 0.2925 | 0.0121 |

Table 5: Ablation study on the sampling steps used for TTC on our long-video benchmark.

subjects that are not specified by the instruction. This indicates that accurately associating indexed references with their intended targets remains challenging in ambiguous scenes.

Second, background replacement may introduce undesired blending between the foreground subjects and the generated background. In such cases, subject boundaries may become less distinct, resulting in local artifacts or inconsistent appearance around the subject-background interface.

Third, the model is less robust to videos containing shot transitions. Across different shots, the correspondence of subjects may not be consistently maintained, causing identity or appearance inconsistencies after a cut. Background replacement may also fail to propagate correctly across shot boundaries, resulting in inconsistent backgrounds between consecutive shots.

Finally, although Long-Horizon Error Control (LHEC) effectively mitigates error accumulation during long-video generation, residual temporal drift may still occur over extended sequences. Fine-grained appearance details and scene characteristics may gradually deviate from the desired result as the generation horizon increases.