STREAMINGR2V: TOWARDS LONG AND INTERACTIVE REFERENCE-TO-VIDEO GENERATION

Anonymous authors
Paper under double-blind review



Figure 1: **StreamingR2V generates long videos with interaction of reference condition**. References can be inserted or removed during autoregressive generation, so subjects enter and leave on cue smoothly without a cut. *Top*: a second subject enters the frame midway. *Middle*: the camera travels with a horse through a city and gathers new subjects into the shot along the way. *Bottom*: a cat exits the frame midway, leaving the ball behind. (*Zoom in for better view*).

ABSTRACT

Reference-to-video (R2V) generation renders user-specified subjects faithfully, but only within a short clip whose reference set is fixed before generation begins. In practice, however, subjects often enter or leave partway through a shot, and users need interactive control over the cast as the video unfolds over a long horizon. We therefore extend R2V to **Streaming Reference-to-Video** (SR2V), which keeps the horizon open and lets references be added or removed as generation proceeds, and identify two challenges that both peak at the moment of interaction. First, history frames carry the identities already rendered, which override the expressiveness of newly inserted references. Second, the chunk where the reference set changes must render a subject unseen in history, which can raise its error, and may accumulate more acutely during generation. We accordingly propose **StreamingR2V** with two complementary designs. The **reference reflective value gate** reconstructs each history token from the references and subtracts the reference-explainable identity from its value, leaving the attention map untouched. The **virtual rollout error chain** adds history-error banks indexed by rollout depth, approximating error accumulation at one forward pass per iteration. On an SR2V benchmark with diverse events of subject entry and exit, StreamingR2V matches or exceeds leading commercial models and several baselines across a wide range of metrics, especially on long-term consistency and interaction compliance. The anonymous demo is available at: `https://sr2v-demo.github.io/`.

1 INTRODUCTION

Reference-to-video (R2V) generation, which synthesizes videos conditioned on user-specified reference images, has become a central downstream application of modern video generation and movie production. Consider the *long take*, a staple of film language in which a scene unfolds without a cut. Its cast rarely stays fixed: characters enter and leave a static frame at scripted moments, or the

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.

camera travels with one subject and gathers others along the way (see Fig. 1). Such a take asks three things of R2V at once: every subject keeps the appearance fixed by its reference image, the shot runs far beyond a few seconds, and, with no cut to hide a seam, it cannot be stitched from separately generated clips. Current R2V models cannot meet these demands together, inheriting two constraints from their video backbones (Wan et al., 2025; Kong et al., 2024): generation is confined to short, fixed-length clips, and the entire reference set must be supplied before generation begins. A reference supplied upfront tends to be rendered from the first frame, so a subject meant to enter later is already in view, and the cast cannot change once generation is underway. As video generation now enters a long, interactive era (Huang et al., 2025a; Yang et al., 2026a; Li et al., 2026a), we argue that R2V urgently needs the same capability: long narrative generation with interactive control over the reference images (e.g., adding or deleting dynamically) as generation proceeds.

We call this capability *Streaming Reference-to-Video* (**SR2V**), a task bridging two lines of work developed in isolation: R2V offers fine-grained control over appearance, but only within a single short clip; Long video generation reaches an open-ended horizon and accepts conditions mid-stream, yet none of them is a reference image pinning down exact appearance. SR2V asks for both: the horizon stays open while the reference set is edited during generation, so subjects enter and leave on cue. It also brings demands neither line is evaluated against: a subject introduced mid-rollout must reach full fidelity at once, its arrival must leave no seam, and every subject must resist drift over the whole horizon. Sitting between the two domains, SR2V invites a first attempt that combines their paradigms: the reference concatenation of R2V (Liu et al., 2025; Deng et al., 2026; Li et al., 2026b) and the practice of conditioning on ground-truth history in long video generation (Yuan et al., 2026b; Wang et al., 2026b). Reference and history latents are concatenated with the target video latents and read by the DiT's full attention, letting the model learn subject and temporal consistency at once.

This paradigm naively accepts a new reference mid-rollout, yet quality collapses once interactive reference set control meets a long rollout due to two challenges. **First**, the subject already rendered into history overrides the reference set. Since the pre-provided subject is usually visible in history, the model need not decide where identity comes from and simply copies it. The shortcut is harmless under a fixed reference set, and fails once a new reference arrives, contaminated by existing identities. The cause is architectural: history and the reference set share one token space and one self-attention, so identity in history cannot be weakened without losing the motion and layout continuity only history provides. **Second**, history is unreliable during rollout, especially under reference interaction. Training draws history from real video whereas inference feeds back the model's own output, so error accumulates chunk by chunk. The accumulation is also uneven under SR2V: at a reference change, the model must render a subject with no evidence in history, which can raise the error of that chunk, which is accumulated autoregressively. Conventional long-video generation settings, which do not involve reference modification, rarely capture this problem. The issue is unavoidable in SR2V, as reference interactions arise during generation and may aggravate error accumulation. In essence, both challenges arise from the interaction between history and newly introduced conditions, but remain separable. The existing subject information carried by history is redundant given a clean reference set, while history can introduce larger errors when new references interact with the rollout. This raises two core questions: **how can we suppress reference-specific information in history, and mitigate error accumulation during a long rollout with the reference interaction?**

To address the above challenges, we propose **StreamingR2V**, an SR2V framework that supports dynamic reference control and enables long, interactive R2V generation. First, to filter redundant reference information entangled in history, we introduce the **Reference Reflective Value Gate**. It projects history and reference values into an subspace, where lightweight attention from history to reference features reconstructs each history token from the references. This reconstruction reveals the components of a history latent that overlap with the reference set, which are then removed by a directed subtraction operator. The module adaptively learns what to remove, without external guidance such as semantic masks. Because it never modifies the attention map, it differs from directly rescaling history attention weights, which trades reference control for temporal continuity. Second, inspired by the error bank design by (Li et al., 2026a; Yuan et al., 2026b; Wang et al., 2026b), we turn its pool of residuals into a dynamic error transition and propose the **Virtual Rollout Error Chain**. The chain deepens as training proceeds: an error injected into the model gives rise to the next error in the chain, which models how error accumulates as the video is generated chunk by chunk. Through these two contributions, StreamingR2V strengthens the expression under dynamic reference control, while mitigating the accumulation of error in history over long generation.

In summary, the contributions of our paper are as follows:

- We propose streaming reference-to-video (SR2V), a task bridging R2V generation and long, interactive video generation, and identify its two challenges: the subjects carried by history override a dynamic reference set, and the error carried by history can grow where the reference set changes.
- We propose two designs that tackle these challenges.Reference reflective value gate adaptively removes the subject information in history that the reference set already explains, keeping a newly inserted subject faithful mid-rollout, while virtual rollout error chain adds error banks indexed by rollout depth, modeling how error accumulates autoregressively without rollout forward passes.
- On an SR2V benchmark covering diverse subject categories and interaction events, StreamingR2V matches or exceeds leading commercial models and achieve state-of-the-art performance especially on long-horizon interactive R2V generation.

2 Related Work

Reference-to-Video Generation. Reference-to-video (R2V) generation synthesizes subject-consistent videos from one or more reference images. Existing methods mainly differ in how the reference is encoded and injected. One line of work (Yuan et al., 2025; He et al., 2024; Chen et al., 2025b; Huang et al., 2025b; Sang et al., 2026; Liang et al., 2025) follows the adapter paradigm from R2V generation, injecting identity features through additional cross-attention. However, aligning identity embeddings with text space or compressing references into semantic tokens loses the high-frequency details needed for fine-grained identity preservation. The second line (Wu et al., 2024; Liu et al., 2025; Cai et al., 2026; Hu et al., 2026; Li et al., 2026b; Deng et al., 2026; 2025; Hu et al., 2025; Chen et al., 2026a; Wang et al., 2026a) encodes references with the video diffusion model's VAE and concatenates the resulting latents with video tokens, preserving visual details without extra modules. Yet both lines assume fixed-length clips: the reference is provided once and remains static, leaving unclear how identity should be re-anchored as generation proceeds chunk by chunk.

Long and Interactive Video Generation. Video foundation models (Wan et al., 2025; Hong et al., 2023; Kong et al., 2024; Seedance et al., 2026) are trained on fixed, short durations, so longer generation faces limited-context inconsistency and rollout error accumulation. To address the former, memory designs compress history or select accessible past frames under a fixed context budget (Henschel et al., 2025; Jin et al., 2025; Zhang et al., 2025; Yuan et al., 2026b; Ji et al., 2025; Chen et al., 2025a; Wang et al., 2026b). To address the latter, forcing strategies convert bidirectional backbones into autoregressive ones and reduce the train-test gap (Gu et al., 2025; Huang et al., 2025a; Chen et al., 2026b; Zhu et al., 2026; Liu et al., 2026; Lu et al., 2026; Yang et al., 2026b), while complementary methods corrupt conditions during training so the model learns to correct historical errors (Li et al., 2026a; Wang et al., 2026b; Po et al., 2026; Guo et al., 2025; Seo et al., 2026). Both directions make history the authoritative evidence for what comes next. Interactive generation breaks this assumption: a mid-generation condition asks the model to override its history. Existing work differs in the incoming signal. Closest to our setting, Yang et al. (2026a); Yesiltepe et al. (2026); Yang et al. (2026b) switch text prompts mid-stream and stabilize transitions; yet text is semantically coarse, so gradual drift may suffice, unlike exact appearance matching. Wang et al. (2026b); Team et al. (2026); Parker-Holder et al. (2025); Zhou et al. (2026) inject geometric control through keyboards or drag vectors, but none supplies a pixel-level visual target. The stronger memory and anti-drift mechanisms anchor generation to history, the more they may resist a reference introduced later.

3 Method

3.1 Preliminaries

Flow matching training with error recycling. We build our method on the flow-matching video foundation model (Wan et al., 2025), which generates in the latent space of a pretrained VAE. Let $\boldsymbol{x}_0$ denote clean latents and $\boldsymbol{x}_1 \sim \mathcal{N}(0, I)$ noise, interpolated by $\boldsymbol{x}_t = (1-t)\boldsymbol{x}_0 + t\boldsymbol{x}_1$. A velocity network v_θ is trained under conditioning $\boldsymbol{c}$ (e.g., prompt, reference images, history frames) by

$$\mathcal{L}_{\text{FM}} \;=\; \mathbb{E}_{t,\boldsymbol{x}_0,\boldsymbol{x}_1,\boldsymbol{c}}\big[\|v_\theta(\boldsymbol{x}_t, t, \boldsymbol{c}) - (\boldsymbol{x}_1 - \boldsymbol{x}_0)\|^2\big], \tag{1}$$

and inference integrates the field from $t{=}1$ to $t{=}0$.

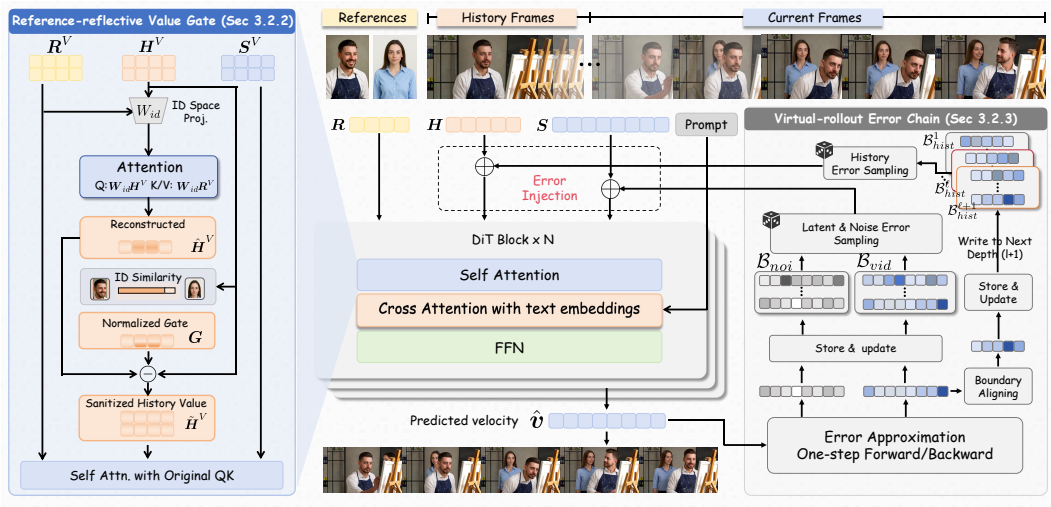

Figure 2: **Overview of StreamingR2V.** StreamingR2V packs reference images, history frames, and current frames into a unified sequence, enabling sufficient interaction through the full attention. The reference reflective value gate is designed to remove redundant reference information in the history frames, thereby enhancing the representation of newly introduced references during interaction control. The virtual rollout error bank is a novel error-recycling strategy tailored for rollout generation, particularly when new references are introduced, where errors are more likely to accumulate.

In long video generation, since training conditions on ground-truth latents while inference conditions on the model's own outputs, errors accumulate across chunks. Stable Video Infinity (SVI) (Li et al., 2026a) mitigates this gap by corrupting the video and conditions during training. From a single forward pass at $\boldsymbol{x}_t$, one velocity estimate $\boldsymbol{v}$ is extrapolated in both directions to approximate the clean latent and the noise, $\hat{\boldsymbol{x}}_0 = \boldsymbol{x}_t - t\,\boldsymbol{v}$ and $\hat{\boldsymbol{x}}_1 = \boldsymbol{x}_t + (1-t)\,\boldsymbol{v}$, yielding a video error and a noise error against their targets. The video error is banked as $\mathcal{B}_{vid} = \{\boldsymbol{\epsilon}_i\}$, where $\boldsymbol{\epsilon}_i = \hat{\boldsymbol{x}}_0^{(i)} - \boldsymbol{x}_0^{(i)}$, so each residual lives in the same space as the signal it will perturb, and is sampled and injected into the video and conditions of the future training data. This is a lightweight add-on: it leaves the architecture and Eq.1 untouched. However, while $\mathcal{B}_{vid}$ is organized by denoising timestep and refreshed as training proceeds, it is *not indexed by rollout depth*: a residual is written back into the same bank from which its own corruption was drawn, so residuals produced under clean and already-perturbed conditions are stored indistinguishably. Therefore, such a plain error arrangement strategy remains challenging for modeling error accumulation during autoregressive long video generation, especially when interactive conditions such as reference images are introduced.

3.2 StreamingR2V

StreamingR2V is designed to autoregressively generate long, subject-consistent videos, where reference conditions can be interactively added or removed to control subject entry or exit in upcoming videos. In Sec. 3.2.1, we first outline our conditioning framework, which obtains the reference and history in a unified manner. We then present the **Reference Reflective Value Gate** in Sec. 3.2.2, which compresses redundant reference information embedded in the history, thereby improving the representation of the newly introduced subjects. Next, we introduce the **Virtual Rollout Error Chain** in Sec. 3.2.3, which models error transitions in chunk-wise autoregressive generation and can improve video consistency at critical points where new interactive references are introduced. The general framework is demonstrated in Fig. 2.

3.2.1 Unified Reference-history Conditioning

An R2V training sample includes a video sequence $\mathbb{V} = \{v_1, v_2, \cdots, v_N\}$ and a set of reference images $\mathbb{I}_R = \{I_1, I_2, \cdots, I_{n_r}\}$, which are normalized to the video's resolution by resizing and white-pixel padding. They are then encoded by the VAE encoder together with the video frames. Following previous methods (Seo et al., 2026; Wang et al., 2026b), the video latents are split into two

parts via a history-current separation strategy, resulting in a unified input sequence $\boldsymbol{u} = [\boldsymbol{R} \,\|\, \boldsymbol{H} \,\|\, \boldsymbol{S}]$, where $\boldsymbol{R}$ represents n_r reference image latents, $\boldsymbol{H}$ contains K latent frames, and $\boldsymbol{S}$ contains F latent frames. After patchification, this sequence is fed into the full attention of the DiT to ensure sufficient feature interaction. During training, only $\boldsymbol{S}$ contributes to the flow matching loss; $\boldsymbol{R}$ and $\boldsymbol{H}$ are discarded before the loss is computed. Reference latents are assigned a RoPE index of -1, acting only as a global condition, whereas history and segment latents share one continuous forward index, with the latents in $\boldsymbol{H}$ assigned RoPE indices of $0, \ldots, K-1$ and those in $\boldsymbol{S}$ continuing from K to $K+F-1$, so that the temporal ramp remains unbroken. During autoregressive inference, $\boldsymbol{H}$ is taken as the last K latent frames of the previously generated $\boldsymbol{S}$. Overall, such a unified input paradigm directly bridges long video generation and R2V, and sets up the basis of StreamingR2V.

3.2.2 Reference Reflective Value Gate

The paradigm above nominally supports chunk-wise SR2V, yet it struggles at the boundary of reference interaction. Since an identity already rendered in the history may have drifted in appearance, the history cannot serve as its identity condition, and its reference image must stay in $\boldsymbol{R}$ throughout. This leads to an asymmetry: previous subjects are present in both $\boldsymbol{H}$ and $\boldsymbol{R}$, while a newly inserted one is present only in $\boldsymbol{R}$, which suppresses the representation of new subjects. The history tokens entering self-attention outnumber the reference latent tokens, and the history is also immediately adjacent to $\boldsymbol{S}$ along the temporal axis. Therefore, the evidence that $\boldsymbol{S}$ reads for previous subjects outweighs that for new ones, manifesting as identity leakage and mixing. However, this reference information in $\boldsymbol{H}$ is also *redundant*: $\boldsymbol{R}$ already specifies every subject that should appear without consistency drift. Thus, reference information in $\boldsymbol{H}$ can be removed, allowing the two conditions to serve their respective purposes: $\boldsymbol{H}$ provides subject-irrelevant consistency, such as background and layout, while $\boldsymbol{R}$ exclusively controls the appearance of the subjects and guarantees consistency. Based on this motivation, we propose the Reference Reflective Value Gate (see the left part of Fig.2), a lightweight module that adaptively identifies and suppresses redundant subject representations in $\boldsymbol{H}$, leaving the full-attention computation untouched. It functions with the following steps:

Step I: Reference-reflective history reconstruction. In each self-attention layer within the DiT block, we define a projection matrix $\boldsymbol{W}_{id}$ and project the value matrices calculated from $\boldsymbol{H}$ and $\boldsymbol{R}$ into a subspace representation with the dimension of d_{id}, annotated as $\boldsymbol{W}_{id}\boldsymbol{H}^V$ and $\boldsymbol{W}_{id}\boldsymbol{R}^V$. Within this subspace, we introduce an attention operation between $\boldsymbol{W}_{id}\boldsymbol{H}^V$ and $\boldsymbol{W}_{id}\boldsymbol{R}^V$. Note that the projected features serve only as query and key, while the unprojected $\boldsymbol{R}^V$ is adopted as the value, so that the output stays in the original value space of $\boldsymbol{H}^V$. The output of this operation represents another history value, reconstructed from the information carried by the reference values:

$$\hat{\boldsymbol{H}}^V = \mathrm{softmax}\left(\frac{\langle \boldsymbol{W}_{id}\,\boldsymbol{H}^V,\ \boldsymbol{W}_{id}\,\boldsymbol{R}^V \rangle}{\sqrt{d_{id}}} \right) \boldsymbol{R}^V, \tag{2}$$

where $\hat{\boldsymbol{H}}^V$ represents the part of the history that can be ***reflected*** by the references. It shares the same space as the original H^V, and can therefore be directly subtracted in the subsequent process.

Step II: Redundancy gating and directed subtraction. In this step, we project both the original history value $\boldsymbol{H}^V$ and its reference-reflective reconstruction $\hat{\boldsymbol{H}}^V$ into the identity subspace and compare them by cosine similarity, which defines the redundant gate $\boldsymbol{G}$:

$$\boldsymbol{G} \;=\; \sigma\Big(\boldsymbol{\alpha} \big(\cos\big(\boldsymbol{W}_{id}\,\boldsymbol{H}^V,\ \boldsymbol{W}_{id}\,\hat{\boldsymbol{H}}^V \big) - \boldsymbol{b} \big) \Big) \;\in\; [0,1]^{n_{\text{hist}}}, \tag{3}$$

where n_{hist} is the token number of $\hat{\boldsymbol{H}}^V$, $\boldsymbol{\alpha}$ and $\boldsymbol{b}$ are learnable affine parameters, σ denotes the *sigmoid* function, and $\cos(\cdot,\cdot)$ denotes the row-wise cosine similarity, computed independently for each of the n_{hist} history tokens, so that $\boldsymbol{G}$ is a per-token gate. A high value of $\boldsymbol{G}$ indicates that the token can largely be represented by the reference set and is therefore redundant in subject information. The sanitized history value is then obtained by a directed subtraction:

$$\tilde{\boldsymbol{H}}^V \;=\; \boldsymbol{H}^V \;-\; \boldsymbol{\gamma}\,\boldsymbol{G} \odot \hat{\boldsymbol{H}}^V, \tag{4}$$

where $\odot$ broadcasts the per-token gate over the channel dimension, and $\boldsymbol{\gamma}$ is a zero-initialized learnable parameter that guarantees the initial behavior is equivalent to the original model.

To sum up, the reference reflective value gate does not alter the original full-attention computation; it only adaptively identifies the part of the history value that carries redundant subject information and keeps it from flowing to subsequent modules.

3.2.3 VIRTUAL-ROLLOUT ERROR CHAIN

The reference reflective value gate removes redundant reference information from the history, leaving it as a condition dedicated to general video consistency. However, in SR2V, history stability remains challenging. As the reference condition changes, the visual content may vary substantially, while the history provides no evidence of the incoming references. This can introduce larger errors that propagate sharply rather than transition smoothly, causing abrupt changes in background and subject appearance in subsequent chunks. As mentioned in Sec. 3.1, a bank that is not indexed by rollout depth cannot represent this regime. It is sufficient when the rollout extrapolates naturally, since the mixed stages are broadly similar. At an interaction boundary, however, these stages differ most, and indiscriminate mixing blurs the crucial distinction. To model error accumulation under condition interaction, we propose the virtual rollout error chain, replacing the original error bank with rollout-depth-indexed error banks, as shown on the right of Fig. 2.

Boundary-aligned error tube for history. Based on existing error recycling mechanism, which the autoregressive inter-chunk error is a single residual. Our history condition spans K frames, so what perturbs it is not a single residual but a temporally ordered segment, and we first specify its shape. Since $\boldsymbol{H}^{(n+1)}$ is the tail of the generated frames $\hat{\boldsymbol{S}}^{(n)}$, the error carried by the history is by definition the tail of the segment prediction residual, and needs no separate definition. Writing $\boldsymbol{\epsilon}^{(n)} = \hat{\boldsymbol{S}}^{(n)} - \boldsymbol{S}^{(n)}$ for the residual already computed at every training step, we collect

$$\boldsymbol{E} \;=\; \boldsymbol{\epsilon}^{(n)}_{F-K+1:F} \;=\; \big[\boldsymbol{\epsilon}^{(n)}_{F-K+1}, \ldots, \boldsymbol{\epsilon}^{(n)}_{F}\big], \tag{5}$$

so that each error bank entry of $\boldsymbol{H}$ is a temporally contiguous tube whose length and ordering match the K history frames it will perturb, at no additional cost.

Depth-indexed banks and relay. The error bank is best organized to mirror chunk-wise rollout, in which every error is inherited from the one before it: $0 \to E_1 \to \cdots$. We refer to the number of accumulation steps a tube represents as its **rollout depth** ℓ, and maintain one bank $\mathcal{B}^{\ell}_{hist}$ per depth, capped at $\ell_{\max}$. Obtaining such a chain directly would require rolling out during training and collecting each chunk's error for the corresponding chain, at $O(\ell)$ forward passes, which is impractical for standard single-clip training. Instead, the virtual rollout error chain advances the error by one step per iteration:

$$\boldsymbol{E}_{\text{in}} \sim \mathcal{B}^{\ell}_{hist}, \qquad \hat{\boldsymbol{v}} = v_\theta\big(\cdot\,,\, \boldsymbol{H} + s\,\boldsymbol{E}_{\text{in}}\big), \qquad \mathcal{B}^{\ell+1}_{hist} \;\leftarrow\; \mathcal{B}^{\ell+1}_{hist} \cup \big\{(f(\hat{\boldsymbol{v}}) - \boldsymbol{S})_{F-K+1:F}\big\}, \tag{6}$$

that is, an error tube drawn from $\mathcal{B}^{\ell}_{hist}$ is injected into the history, one standard forward pass produces a new segment, and its boundary-aligned tail is stored one depth deeper in $\mathcal{B}^{\ell+1}_{hist}$. The relay is necessary, not merely convenient. After one-step prediction from velocity denoted as $\hat{\boldsymbol{S}} = f(\hat{\boldsymbol{v}})$, a segment tail lies in the future of the history that produced it, so it cannot perturb that history and must wait for a later iteration. Each bank therefore collects tails from clips hundreds of iterations apart, and no continuous rollout occurs. This is not a problem, because we model the distribution of error states rather than any specific video, aligning with the strategy of SVI. At one forward pass per iteration, the chain approximates with $O(1)$ cost what true rollout attains with $O(\ell)$.

4 EXPERIMENTS

Implementation Details. We implement StreamingR2V on Wan2.1-I2V-14B. We curate our training data from Phantom-Data (Chen et al., 2026c) and OpenS2V-5M (Yuan et al., 2026a), filtering for high aesthetic quality and for samples containing multiple individuals, yielding 110k and 80k 480p clips, respectively. Based on these filtered data, we further extract about 9k clips that contain explicit behavior of entering and exiting the subject. During training, we extract 82 frames per video sample, taking the first $\mathbf{17}$ as history frames and the last $\mathbf{65}$ as the current frames, which accordingly set their latent length $K = 5$ and $F = 17$, respectively. The error chain is capped at $\ell_{\max} = 3$. The model is trained with LoRA with a rank of 128 on 32 A100 GPUs for 3500 steps, and the learning rate is $1e-4$. More details about the training dataset and recipe are available in Appendix A.1.

Table 1: **Quantitative comparison results of SR2V**. The best scores are shown in **bold**, and the second-best are underlined. Video quality reports aesthetic score (Aes), image quality (IQ), motion smoothness (MS), background consistency (BC) and dynamic degree (DD); R2V consistency reports GME Score, Face Similarity (FS) and DINO-I; interaction compliance reports entry success rate (ESR), premature appearance rate (PR) and exit success rate (XSR) and residual rate (RR). "*" marks commercial models, which accept all conditions in a single request.

Method	Video Quality					R2V Consistency			Interaction Compliance			
	Aes ↑	IQ ↑	MS↑	BC↑	DD↑	GME↑	FS ↑	DINO-I↑	ESR↑	PR↓	XSR↑	RR ↓
Evaluation horizon of 200 frames												
Kling-Omni (*) (Kling, 2026)	0.595	0.723	0.992	0.904	0.474	0.608	0.588	0.625	0.971	0.675	**0.929**	0.121
MiniMax-H3 (*) (MiniMax, 2026)	**0.606**	**0.731**	**0.994**	0.931	**0.533**	0.601	0.592	0.614	**1.000**	0.739	0.642	0.226
VACE (Jiang et al., 2025) + SVI	0.570	0.711	0.989	0.925	0.475	0.573	0.425	0.641	0.941	0.579	0.786	0.198
Phantom (Liu et al., 2025) + SVI	0.564	0.708	0.988	0.902	0.463	0.559	0.576	0.643	0.794	0.081	**0.929**	0.206
MAGREF (Deng et al., 2026) + SVI	0.566	0.702	0.990	0.923	0.465	0.578	0.446	0.634	0.794	0.102	0.857	0.231
RefAlign (Wang et al., 2026a) + SVI	0.577	0.718	0.991	0.929	0.465	0.578	0.532	0.640	0.882	0.085	0.714	0.342
SVI+R2V SFT	0.571	0.715	0.990	0.919	0.483	0.575	0.513	0.632	0.735	0.129	0.857	0.153
StreamingR2V (Ours)	0.601	0.728	**0.994**	**0.942**	0.523	**0.615**	**0.596**	**0.669**	**1.000**	**0.023**	**0.929**	**0.094**
Evaluation at the extended horizon over 200 frames												
VACE (Jiang et al., 2025) + SVI	0.560	0.720	0.988	0.869	0.274	0.565	0.399	0.604	0.824	0.139	0.907	0.182
Phantom (Liu et al., 2025) + SVI	0.549	0.718	0.991	0.876	0.375	0.561	0.470	0.564	0.882	0.094	0.881	0.194
MAGREF (Deng et al., 2026) + SVI	0.562	0.715	0.991	0.867	0.338	0.562	0.386	0.578	0.765	0.083	0.907	0.148
RefAlign (Wang et al., 2026a) + SVI	0.569	0.720	0.992	0.884	0.319	0.571	0.490	0.597	0.778	0.126	0.837	0.165
SVI+R2V SFT	0.558	0.723	0.992	0.908	0.398	0.543	0.492	0.612	0.765	0.101	0.884	0.134
StreamingR2V (Ours)	**0.578**	**0.736**	**0.993**	**0.918**	**0.474**	**0.602**	**0.545**	**0.645**	**0.941**	**0.044**	**0.930**	**0.103**

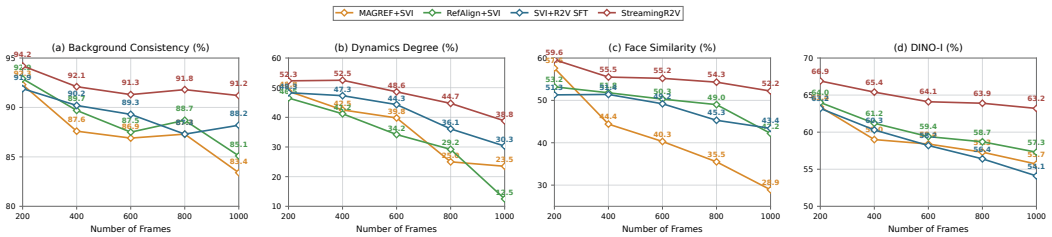

Figure 3: Robustness comparison on different metrics across different video lengths.

Evaluation. As SR2V has no public evaluation protocol, we curate our own benchmark. Reference images are drawn from public reference-to-image/video benchmarks (Yuan et al., 2026a; Ruiz et al., 2023), covering diverse categories, including humans, animals, products, and daily objects. We generate prompts with reference interaction using Qwen3 (Yang et al., 2025). In total, we curate **120** evaluation samples, ranging from **200** frames to **1200** frames. We report metrics in 3 categories. (1) ***Video quality***, quantified by aesthetic score, image quality, motion smoothness, background consistency, and dynamics degree, from VBench (Huang et al., 2024). (2) ***R2V Consistency***, which includes the GME score for prompt following and face similarity from OpenS2V-Eval (Yuan et al., 2026a). For general objects, we calculate the DINO-I (Oquab et al., 2023). To better evaluate the subject consistency around the interaction, we report Face Similarity and DINO-I scores before and after the interaction cue across chunks, noted as $\mathrm{FS}_{b/a}$ and $\mathrm{DINO\text{-}I}_{b/a}$ respectively. (3) ***Interaction compliance***. To verify that entering subjects appear faithfully only after the interaction, rather than in advance, and that exiting subjects disappear from the scene rather than remain present, we propose four types of metrics, including entry successful rate (ESR), premature rate (PR), exit successful rate (XSR), and residual rate (RR). Details of these metrics are in Appendix A.2.

Baselines. To the best of our knowledge, SR2V, targeting long-video generation with interactive reference conditions, is a relatively new task, and therefore directly comparable baselines hardly exist. We set up the comparison with three types of baselines: (1) ***Commercial models***, which can generate longer videos than Wan2.1 but still receive all conditions up front; (2) ***Two-stage baselines***, which re-invoke a short-clip R2V model when the reference set changes and extend each generated chunk with a long-video generator until the next change. For this setting, we choose competitive R2V models including VACE(Jiang et al., 2025), Phantom(Liu et al., 2025), MAGREF(Deng et al., 2026) and RefAlign (Wang et al., 2026a) and using SVI as the long-video generator. All models used in these two stages are tuned from Wan2.1-14B to ensure comparable foundation-model capability; (3) ***A naive baseline***, denoted as **SVI+R2V SFT**, is built upon SVI itself and trained on the same dataset using the original error bank. Since it is jointly trained with both history and reference conditions, it acquires SR2V capability and is therefore directly comparable to StreamingR2V.

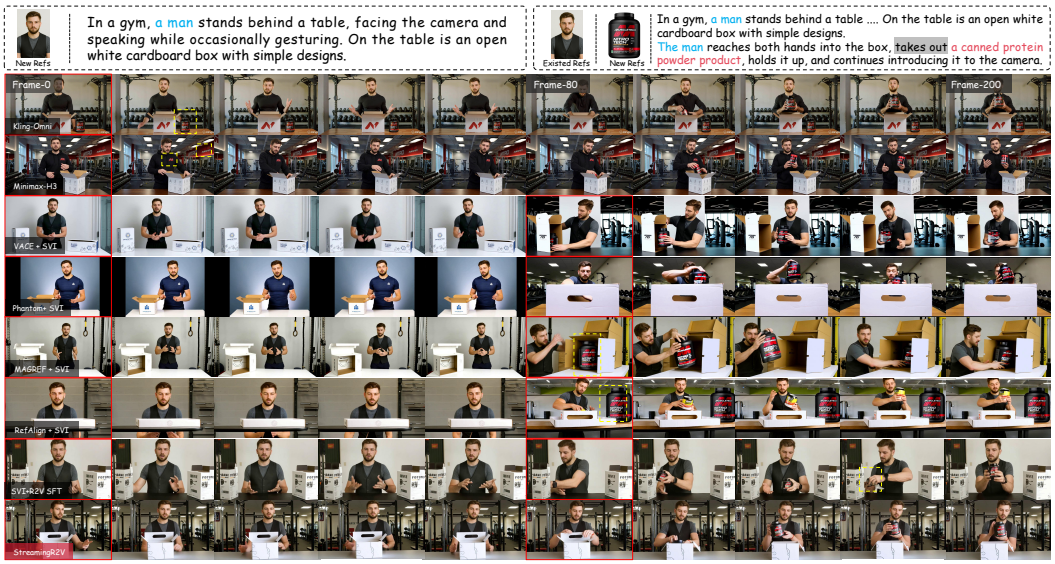

Figure 4: **Qualitative comparison of the SR2V task on 200-frame horizons.** The initial frames for each interactive condition setting are highlighted with the red box. *Zoom in for the best view.*



Figure 5: **Qualitative comparison on longer horizons with complex SR2V interaction patterns.** When multiple subject insertions are included, StreamingR2V demonstrates the best performance in quality stability and faithful interaction patterns between subjects. *Zoom in for the best view.*

4.1 Main Results

Quantitative Comparison. At the 200-frame SR2V evaluation setting (Tab. 1 top), StreamingR2V matches or exceeds R2V specialists on the core identity metrics, confirming that its long-horizon and interactive capabilities come at no cost in short-range generation. From 400 to 1200 frames (Tab. 1 bottom), StreamingR2V leads on all core metrics, especially background consistency, subject consistency, and prompt following. On the critical metrics related to interactive compliance, StreamingR2V also demonstrates notable advantages. Tab. 2 isolates the chunks around reference interactions, where StreamingR2V achieves the best consistency for the inserted subject and the least degradation of those already present. Commercial models have significantly higher PR, which can be attributed to the up-front injection of all references. Furthermore, Fig. 3 plots several metrics against frame number, and our curves are the flattest, indicating the least error accumulation in long-horizon SR2V generation. Overall, StreamingR2V demonstrates the most competitive performance across several metrics and shows robust performance across video generation of different lengths.

Qualitative Comparison. Fig. 4 compares 200-frame results in a scene where *a man picks up a product from a box*. Because commercial models receive all conditions upfront, the product or its parts appear before they should emerge from the box. Such leakage results in their higher PR. The two-stage baselines place the product at the correct time but cause background shifts at the interaction boundary, rendering the product abruptly rather than introducing it smoothly. StreamingR2V shows neither issue. Fig. 5 considers a more challenging long-horizon setting, where inserted references must appear and integrate into the ongoing action, as in *the woman keeps playing the guitar*. The prompt also specifies camera movement, which remains continuous during the reference interaction. StreamingR2V consistently performs a smooth appearance of new references and the most faithful interaction motions. More qualitative comparisons are provided in Appendix A.3.

Table 2: Quantitative results of SR2V consistency on chunks before/after interaction cue.

Method	MS↑	BC↑	FS_b ↑	FS_a↑	DINO-I_b ↑	DINO-I_a↑
VACE + SVI	0.988	0.887	0.556	0.315	0.683	0.535
Phantom + SVI	0.989	0.888	0.568	0.445	0.671	0.558
MAGREF + SVI	0.989	0.880	0.517	0.357	0.655	0.570
RefAlign + SVI	0.990	0.887	<u>0.581</u>	<u>0.486</u>	<u>0.701</u>	<u>0.573</u>
SVI+R2V SFT	<u>0.991</u>	<u>0.912</u>	0.494	0.377	0.668	0.483
StreamingR2V	**0.993**	**0.925**	**0.601**	**0.533**	**0.713**	**0.632**

Table 3: Ablation study on core designs. Plain setting is equivalent to SVI+R2V SFT.

Setting		Consistency Metrics					
Gate	Chain	MS↑	BC↑	FS_b↑	FS_a↑	DINO-I_b ↑	DINO-I_a↑
✗	✗	0.991	0.912	0.494	0.377	0.668	0.483
✗	✓	0.992	0.916	0.528	0.414	0.680	0.515
✓	✗	0.993	0.913	0.555	0.493	0.681	0.614
✓	✓	**0.993**	**0.925**	**0.601**	**0.533**	**0.713**	**0.632**

4.2 Ablation Study

Since both of our designs focus on improving the SR2V performance at the interaction point, we ablate our designs and evaluate them following the metrics in Tab. 2. Quantitative and qualitative results are shown in Tab. 3 and Fig. 6. *Gate* and *Chain* are short for reference reflective value gate and virtual rollout error chain. When both contributions are ablated, the model degenerates to SVI+R2V SFT.



Figure 6: Qualitative ablation studies.

Reference reflective value gate. Rows (1, 3) and (2, 4) of Tab. 3 and Fig. 6 show that the gaps in Face-Sim and DINO-I between the chunks before and after an interaction shrink substantially once the reference reflective value gate is applied. Without it, the newly inserted reference is reproduced with low fidelity or even blends into an existing identity. Scene consistency across the interaction, however, remains challenging for this design alone, since both Row 1 and 3 still demonstrate the background inconsistency.

Virtual rollout error chain. Rows (1, 2) and (3, 4) of Tab. 3 and Fig. 6 show that the virtual rollout error chain improves both subject and background consistency, although its gain in subject fidelity is smaller than that brought by the reference reflective value gate.

The two designs are therefore complementary: together they improve subject consistency as well as general background and motion consistency. Additional ablations are provided in Appendix. A.3.2.

5 Conclusion

We introduced *streaming reference-to-video* (SR2V), a novel task that generates long subject-consistent video with reference interaction: the horizon stays open and references can be added or removed as generation proceeds, so subjects enter and leave a long take on cue. Interaction raises two challenges: the identity already in history overrides a newly inserted reference, and the chunk where the reference set changes becomes the least reliable history for those that follow. StreamingR2V addresses them with the reference reflective value gate, which strips reference-explainable identity from history, and the virtual rollout error chain, which organizes recycled errors by rollout depth. On our SR2V benchmark, it matches or exceeds leading commercial models on video quality and leads all baselines on core metrics of R2V consistency and interaction compliance. We hope SR2V encourages further work on customized generation that stays controllable as video unfolds.

AI USE STATEMENT

The authors used generative AI tools to refine the manuscript's language, improving clarity, grammar, and conciseness. The authors carefully reviewed and revised all AI-assisted edits to ensure that they accurately reflected the intended scientific meaning. The authors take full responsibility for the final content, claims, and conclusions of this work.

ETHICS STATEMENT

Reference images used in the experiments are sourced from public datasets and benchmarks (see Appendix. A.2 for details), thereby reducing the need to collect personal images from individuals. We take care to use these materials only for research and evaluation purposes, in accordance with applicable dataset terms and data licenses. We support the responsible use of the proposed method, with due respect for privacy, consent, and relevant legal or ethical requirements. We discourage any use of this technology for deceptive, misleading, non-consensual, or otherwise harmful content.

REPRODUCIBILITY STATEMENT

We describe the core architecture and training objectives of StreamingR2V in Sec. 3. The data sources and processing procedures are detailed in Appendix A.1, where we also provide the training recipe, detailed parameter settings, and inference configuration. Additional evaluation protocols are provided in Appendix A.2.

REFERENCES

Yuanhao Cai, He Zhang, Xi Chen, Jinbo Xing, Yiwei Hu, Yuqian Zhou, Kai Zhang, Zhifei Zhang, Soo Ye Kim, Tianyu Wang, et al. Omnivcus: Feedforward subject-driven video customization with multimodal control conditions. *Advances in Neural Information Processing Systems*, 38: 115404–115423, 2026.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *International conference on learning representations*, volume 2026, pp. 138846–138923, 2026.

Liyang Chen, Tianxiang Ma, Jiawei Liu, Bingchuan Li, Zhuowei Chen, Lijie Liu, Xu He, Gen Li, Qian He, and Zhiyong Wu. Human-centric video generation via collaborative multi-modal conditioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 2939–2947, 2026a.

Nan Chen, Mengqi Huang, Yihao Meng, and Zhendong Mao. Longanimation: Long animation generation with dynamic global-local memory. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10032–10042. IEEE, 2025a.

Shuo Chen, Cong Wei, Sun Sun, Ping Nie, Kai Zhou, Ge Zhang, Ming-Hsuan Yang, and Wenhu Chen. Context forcing: Consistent autoregressive video generation with long context. *arXiv preprint arXiv:2602.06028*, 2026b.

Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Yuwei Fang, Kwot Sin Lee, Ivan Skorokhodov, Kfir Aberman, Jun-Yan Zhu, Ming-Hsuan Yang, and Sergey Tulyakov. Multi-subject open-set personalization in video generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6099–6110. IEEE, 2025b.

Zhuowei Chen, Bingchuan Li, Tianxiang Ma, Lijie Liu, Mingcong Liu, Yunsheng Jiang, Li Li, Xinghui Li, Liyang Chen, Siyu Zhou, et al. Phantom-data: Towards a general subject-consistent video generation dataset. In *International Conference on Learning Representations*, volume 2026, pp. 73018–73037, 2026c.

Yufan Deng, Xun Guo, Yizhi Wang, Jacob Zhiyuan Fang, Angtian Wang, Shenghai Yuan, Yiding Yang, Bo Liu, Haibin Huang, and Chongyang Ma. Cinema: Coherent multi-subject video generation via mllm-based guidance. *arXiv preprint arXiv:2503.10391*, 2025.

Yufan Deng, Yuanyang Yin, Xun Guo, Yizhi Wang, Jacob Zhiyuan Fang, Shenghai Yuan, Yiding Yang, Angtian Wang, Bo Liu, Haibin Huang, and Chongyang Ma. MAGREF: Masked guidance for any-reference video generation with subject disentanglement. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Nbl43eAVaE.

Yuchao Gu, Weijia Mao, and Mike Zheng Shou. Long-context autoregressive video modeling with next-frame prediction. *arXiv preprint arXiv:2503.19325*, 2025.

Yuwei Guo, Ceyuan Yang, Hao He, Yang Zhao, Meng Wei, Zhenheng Yang, Weilin Huang, and Dahua Lin. End-to-end training for autoregressive video diffusion via self-resampling. *arXiv preprint arXiv:2512.15702*, 2025.

Xuanhua He, Quande Liu, Shengju Qian, Xin Wang, Tao Hu, Ke Cao, Keyu Yan, and Jie Zhang. Id-animator: Zero-shot identity-preserving human video generation. *arXiv preprint arXiv:2404.15275*, 2024.

Roberto Henschel, Levon Khachatryan, Hayk Poghosyan, Daniil Hayrapetyan, Vahram Tadevosyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Streamingt2v: Consistent, dynamic, and extendable long video generation from text. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2568–2577. IEEE, 2025.

Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=rB6TpjAuSRy.

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*, 2025.

Teng Hu, Zhentao Yu, Zhengguang Zhou, Jiangning Zhang, Yuan Zhou, Qinglin Lu, and Ran Yi. Polyvivid: Vivid multi-subject video generation with cross-modal interaction and enhancement. *Advances in Neural Information Processing Systems*, 38:49394–49420, 2026.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *Advances in Neural Information Processing Systems*, 38:167283–167308, 2025a.

Yuzhou Huang, Ziyang Yuan, Quande Liu, Qiulin Wang, Xintao Wang, Ruimao Zhang, Pengfei Wan, Di Zhang, and Kun Gai. Conceptmaster: Multi-concept video customization on diffusion transformer models without test-time tuning. *arXiv preprint arXiv:2501.04698*, 2025b.

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.

Sihui Ji, Xi Chen, Shuai Yang, Xin Tao, Pengfei Wan, and Hengshuang Zhao. Memflow: Flowing adaptive memory for consistent and efficient long video narratives. *arXiv preprint arXiv:2512.14699*, 2025.

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

Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. In *International Conference on Learning Representations*, volume 2025, pp. 23378–23402, 2025.

Kling. Kling api. https://klingai.com/global/, 2026. Accessed: 2026-09-01.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Wuyang Li, Wentao Pan, Po-Chien Luan, Yang Gao, and Alexandre Alahi. Stable video infinity: Infinite-length video generation with error recycling. In *International Conference on Learning Representations*, volume 2026, pp. 23406–23432, 2026a.

Zhaoyang Li, Dongjun Qian, Kai Su, Xiangyang Xia, Chang Liu, Wenfei Yang, Tianzhu Zhang, Zehuan Yuan, et al. Bindweave: Subject-consistent video generation via cross-modal integration. In *International Conference on Learning Representations*, volume 2026, pp. 107960–107979, 2026b.

Feng Liang, Haoyu Ma, Zecheng He, Tingbo Hou, Ji Hou, Kunpeng Li, Xiaoliang Dai, Felix Juefei-Xu, Samaneh Azadi, Animesh Sinha, et al. Movie weaver: Tuning-free multi-concept video personalization with anchored prompts. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13146–13156. IEEE, 2025.

Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. In *International Conference on Learning Representations*, volume 2026, pp. 91177–91196, 2026.

Lijie Liu, Tianxiang Ma, Bingchuan Li, Zhuowei Chen, Jiawei Liu, Gen Li, Siyu Zhou, Qian He, and Xinglong Wu. Phantom: Subject-consistent video generation via cross-modal alignment. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14951–14961. IEEE, 2025.

Yunhong Lu, Yanhong Zeng, Haobo Li, Hao Ouyang, Qiuyu Wang, Ka Leong Cheng, Jiapeng Zhu, Hengyuan Cao, Zhipeng Zhang, Xing Zhu, et al. Reward forcing: Efficient streaming video generation with rewarded distribution matching distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 34385–34397, 2026.

MiniMax. Minimax-h3. `https://www.minimax.io/`, 2026. Accessed: 2026-09-05.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Jack Parker-Holder, Shlomi Fruchter, et al. Genie 3: A new frontier for world models. *Google DeepMind Blog*, 2025.

Ryan Po, Eric Ryan Chan, Changan Chen, and Gordon Wetzstein. Bagger: Backwards aggregation for mitigating drift in autoregressive video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 43727–43739, 2026.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22500–22510. IEEE, 2023.

Shen Sang, Tiancheng Zhi, Tianpei Gu, Jing Liu, and Linjie Luo. Lynx: Towards high-fidelity personalized video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9192–9202, 2026.

Team Seedance, De Chen, Liyang Chen, Xin Chen, Ying Chen, Zhuo Chen, Zhuowei Chen, Feng Cheng, Tianheng Cheng, Yufeng Cheng, et al. Seedance 2.0: Advancing video generation for world complexity. *arXiv preprint arXiv:2604.14148*, 2026.

Junyoung Seo, Hyunwook Choi, Minkyung Kwon, Jinhyeok Choi, Siyoon Jin, Gayoung Lee, Junho Kim, JoungBin Lee, Geonmo Gu, Dongyoon Han, et al. Grounding world simulation models in a real-world metropolis. *arXiv preprint arXiv:2603.15583*, 2026.

Robbyant Team, Zelin Gao, Qiuyu Wang, Yanhong Zeng, Jiapeng Zhu, Ka Leong Cheng, Yixuan Li, Hanlin Wang, Yinghao Xu, Shuailei Ma, et al. Advancing open-source world models. *arXiv preprint arXiv:2601.20540*, 2026.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Lei Wang, YuXin Song, Ge Wu, Haocheng Feng, Hang Zhou, Jingdong Wang, Yaxing Wang, and Jian Yang. Refalign: Representation alignment for reference-to-video generation. *arXiv preprint arXiv:2603.25743*, 2026a.

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*, 2026b.

Tao Wu, Yong Zhang, Xiaodong Cun, Zhongang Qi, Junfu Pu, Huanzhang Dou, Guangcong Zheng, Ying Shan, and Xi Li. Videomaker: Zero-shot customized video generation with the inherent force of video diffusion models. *arXiv preprint arXiv:2412.19645*, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Ying-Cong Chen, Yao Lu, Song Han, and Yukang Chen. Longlive: Real-time interactive long video generation. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=nCAODkpsPJ.

Yang Yang, Tianyi Zhang, Wei Huang, Jinwei Chen, Boxi Wu, Xiaofei He, Deng Cai, Bo Li, and Peng-Tao Jiang. Anchor forcing: Anchor memory and tri-region rope for interactive streaming video diffusion. *arXiv preprint arXiv:2603.13405*, 2026b.

Hidir Yesiltepe, Tuna Meral, Adil Kaan Akan, Kaan Oktay, and Pinar Yanardag. Infinity-rope: Action-controllable infinite video generation emerges from autoregressive self-rollout. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 40256–40265, 2026.

Shenghai Yuan, Jinfa Huang, Xianyi He, Yunyang Ge, Yujun Shi, Liuhan Chen, Jiebo Luo, and Li Yuan. Identity-preserving text-to-video generation by frequency decomposition. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12978–12988. IEEE, 2025.

Shenghai Yuan, Xianyi He, Yufan Deng, Yang Ye, Jinfa Huang, Chongyang Ma, Jiebo Luo, Li Yuan, et al. Opens2v-nexus: A detailed benchmark and million-scale dataset for subject-to-video generation. *Advances in Neural Information Processing Systems*, 38, 2026a.

Shenghai Yuan, Yuanyang Yin, Zongjian Li, Xinwei Huang, Xiao Yang, and Li Yuan. Helios: Real real-time long video generation model. *arXiv preprint arXiv:2603.04379*, 2026b.

Lvmin Zhang, Shengqu Cai, Muyang Li, Gordon Wetzstein, and Maneesh Agrawala. Frame context packing and drift prevention in next-frame-prediction video diffusion models. *Advances in Neural Information Processing Systems*, 38:30546–30566, 2025.

Junbao Zhou, Yuan Zhou, Kesen Zhao, Qingshan Xu, Beier Zhu, Richang Hong, and Hanwang Zhang. Streaming drag-oriented interactive video manipulation: Drag anything, anytime! In *International Conference on Learning Representations*, volume 2026, pp. 129352–129378, 2026.

Hongzhou Zhu, Min Zhao, Guande He, Hang Su, Chongxuan Li, and Jun Zhu. Causal forcing: Autoregressive diffusion distillation done right for high-quality real-time interactive video generation. *arXiv preprint arXiv:2602.02214*, 2026.

A APPENDIX

Our Appendix consists of the following parts:

- Appendix. A.1 details our implementation, including the curation pipeline for the SR2V training data, the full training recipe, and the inference settings.
- Appendix. A.2 presents the evaluation details, including the construction of the evaluation datasets, the evaluation metrics, and an introduction to comparable baselines.
- Appendix.. A.3 provides additional qualitative comparisons with baselines, as well as ablation studies under more detailed parameter settings, including the dimension of the ID subspace and the choice of the error-chain cap depth ℓ_{max}. We also present a runtime analysis of the reference-reflective value gate and a human preference study.
- Appendix. A.4 showcases additional generation results for StreamingR2V, together with examples of practical applications.
- Appendix. A.5 discusses potential application scenarios for StreamingR2V, along with its limitations and directions for future work.

The original video samples for all examples mentioned in both the main submission and the appendix are provided on our demo webpage: `https://sr2v-demo.github.io/`. We strongly recommend viewing them for a direct and clear demonstration of StreamingR2V's capabilities and its improvements over previous methods.

A.1 IMPLEMENTATION DETAILS

A.1.1 SR2V DATASET CURATION

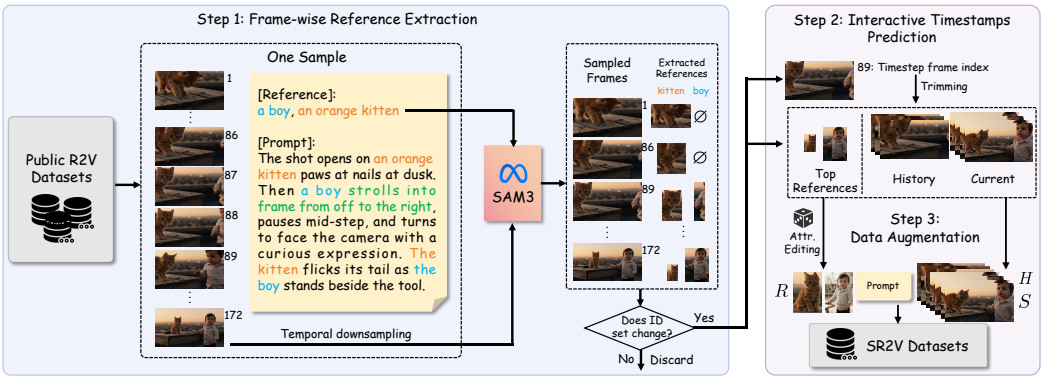

Figure 7: **Overview of SR2V dataset curation pipeline.** Given a standard R2V data sample containing an original video, a prompt, and reference images, the pipeline first applies an instance-level subject segmentation model to video frames to extract frame-wise reference results. By tracking how these reference results change along the frame sequence, the pipeline determines whether a subject entering or exiting event occurs. If such a change is detected, the timestamp at which the reference result changes is regarded as the critical timestamp of the entering/exiting behavior. Finally, according to the training configuration, the frames before this timestamp are used as history frames, while the frames after it are used as current frames.

Public R2V datasets (Yuan et al., 2026a; Chen et al., 2026c) do contain clips where a subject enters or exits, but these events are scattered at arbitrary positions in videos across different lengths, and random sampling rarely places one at the boundary between history and current frames, which is exactly where StreamingR2V needs supervision. We therefore curate such clips explicitly, as illustrated in Fig. 7. We run Segment-Anything-3 (Carion et al., 2026) over each source video to obtain frame-wise reference sets, and keep only the samples whose sets change along the sequence, taking the change point as the history-current boundary. Reference images are then taken from the highest-confidence segmentation of each subject. To mitigate the copy-and-paste effect, an editing model further perturbs non-identity attributes such as viewing angle and lighting with a given probability.

Table 4: Training hyperparameters of StreamingR2V.

Group	Hyperparameter	Value
Optimization	Trainable parameters	LoRA of DiT, $\boldsymbol{W}_{id}$, $\boldsymbol{\alpha}$, $\boldsymbol{b}$, $\boldsymbol{\gamma}$.
	LoRA rank / $alpha$	128 / 128
	LoRA initialization	Kaiming
	LoRA position	`q,k,v,o,ffn.0,ffn.2`
	Optimizer	AdamW
	Learning rate	$1e-4$
	(β_1, β_2)	$(0.9, 0.999)$
	Gradient checkpointing	enabled (with offload)
Data & schedule	Resolution	832×480
	Frames per sample	82
	Max reference images	5
	Training steps	3500
	GPUs / batch size per GPU	$32 \times$ A100 / 1
Video & noise error bank	Latent injection probability	0.9
	Noise injection probability	0.01
	Capacity per timestep grid	500
History virtual rollout error chain	Max rollout depth $\ell_{\max}$	3
	History injection scale s	1.0
	History injection probability	0.9
	Capacity per timestep grid	500
Value gate	ID subspace dimension d_{id}	64
	$\boldsymbol{\gamma}$ initialization	0

To sum up, the final training dataset consists of three parts: **110k** filtered Phantom-Data clips, **80k** filtered OpenS2V clips, and an additional **9k** processed SR2V clips. The SR2V clips are standardized to 82 frames and feature explicit subject entry and exit behaviors. All three subsets are mixed during the 3,500-step training process.

A.1.2 Training Details

For the detailed supplement to implementation details mentioned in Sec. 4, Tab. 4 lists the full training recipe. We train a LoRA adapter for the DiT Blocks rather than the full backbone, and the reference reflective value gate is learned alongside a lightweight adaptation of the base model. The gate's own parameters, $\boldsymbol{W}_{id}$, $\boldsymbol{\alpha}$, $\boldsymbol{b}$ and $\boldsymbol{\gamma}$, are trained in full precision, with $\boldsymbol{\gamma} = 0$ at initialization so that the module starts as an identity mapping.

Our recipe keeps the two error banks $\mathcal{B}_{vid}$ and $\mathcal{B}_{noi}$ of SVI, holding video-latent and noise residuals, with their original sampling and refresh procedure as well as capacity. Both describe the error of a single denoising pass; since StreamingR2V denoises each chunk in the same way, this part carries over unchanged. What they leave open is the question of how the error of one chunk, once it becomes the history of the next, shapes the error of that chunk. The virtual rollout error chain addresses precisely this, corrupting the history while the SVI banks corrupt the latents and the noise, so the two do not conflict.

Last but not least, during inference for the first chunk, no historical context is available. To reflect this setting during training, we drop the history condition with a probability of 0.5 and train the model using only the reference images and prompts as conditions.

Alg. 1 states one iteration with virtual rollout error chain, where the branch on p_{inject} is what makes the depth index meaningful. An iteration conditioned on clean history stands at depth 0, so its error has accumulated one step and seeds $\mathcal{B}^1_{hist}$; one drawing $\boldsymbol{E}_{\text{in}}$ from $\mathcal{B}^{\ell}_{hist}$ produces an error one step deeper and writes to $\mathcal{B}^{\ell+1}_{hist}$. Sampling is restricted to non-empty banks with $\ell \leq \ell_{\max}$, so the chain deepens only as fast as the banks below it fill, and with $\ell_{\max} = 3$ the deepest injected tube.

Algorithm 1 Virtual rollout error chain: one training iteration (when history condition is included)

Require: depth-indexed banks $\{\mathcal{B}^1_{hist}, \ldots, \mathcal{B}^{\ell_{\max}}_{hist}\}$; video latent & noise error banks $\mathcal{B}_{vid}, \mathcal{B}_{noi}$; history injection probability p_{inject}; injection scale s; depth cap $\ell_{\max}$

sample an R2V data and form $\boldsymbol{u} = [\boldsymbol{R} \,\|\, \boldsymbol{H} \,\|\, \boldsymbol{S}]$ ▷ Sec. 3.2.1
$\mathcal{A} \leftarrow \{\, \ell \le \ell_{\max} \;:\; \mathcal{B}_\ell \neq \emptyset \,\}$ ▷ banks available for sampling
if $\text{rand}() > p_{\text{inject}}$ **or** $\mathcal{A} = \emptyset$ **then**
  $\ell_{\text{in}} \leftarrow 0, \quad \boldsymbol{E}_{\text{in}} \leftarrow \boldsymbol{0}$ ▷ clean history: the chain starts here
else
  $\ell_{\text{in}} \sim \mathcal{U}(\mathcal{A}), \quad \boldsymbol{E}_{\text{in}} \sim \mathcal{U}(\mathcal{B}^{\ell_{\text{in}}}_{hist})$ ▷ history already ℓ_{in} steps deep
end if
$\boldsymbol{H} \leftarrow \boldsymbol{H} + s\,\boldsymbol{E}_{\text{in}}$ ▷ corrupt the history condition
$\boldsymbol{\epsilon}_{vid} \sim \mathcal{B}_{vid}, \quad \boldsymbol{\epsilon}_{noi} \sim \mathcal{B}_{noi}, \quad t \sim \mathcal{U}(0,1), \quad \boldsymbol{x}_1 \sim \mathcal{N}(0, I)$
$\boldsymbol{S}_t \leftarrow (1-t)\,(\boldsymbol{S} + s\,\boldsymbol{\epsilon}_{vid}) + t\,(\boldsymbol{x}_1 + s\,\boldsymbol{\epsilon}_{noi})$ ▷ Naive latent and noise corruption
$\hat{\boldsymbol{v}} \leftarrow v_\theta\big([\boldsymbol{R} \,\|\, \tilde{\boldsymbol{H}} \,\|\, \boldsymbol{S}_t]\big)$ ▷ one standard forward pass
update θ with $\mathcal{L}_{\text{FM}}$ on $\boldsymbol{S}$ ▷ Eq. 1 unchanged
$\boldsymbol{E}_{\text{out}} \leftarrow (f(\hat{\boldsymbol{v}}) - \boldsymbol{S})_{F-K+1:F}$ ▷ boundary-aligned tail after prediction approximation, Eq. 6
$\mathcal{B}^{\ell_{\text{in}}+1}_{hist} \leftarrow \mathcal{B}^{\ell_{\text{in}}+1}_{hist} \cup \{\boldsymbol{E}_{\text{out}}\}$, evicting the oldest tube if full
refresh $\mathcal{B}_{vid}, \mathcal{B}_{noi}$ from $\hat{\boldsymbol{S}}$ as in SVI ▷ unchanged from the base recipe

A.1.3 Inference Details.

During inference, we first generate a 65-frame chunk conditioned only on the initial reference images and prompt, using the standard 50 denoising steps. Afterwards, we retain the last 17 frames of the currently generated chunk as history and continue the subsequent autoregressive generation.

A.2 Evaluation Details

SR2V is a relatively new task, and while existing benchmarks provide useful references, directly reusing them is not well suited to our setting for reasons beyond their limited availability. An R2V benchmark often pairs a single fixed reference set with a single prompt over a short clip, so it provides no timeline on which the arrival of a subject could be placed; simply extending its duration does not help, since every reference stays available from the first frame and the entering and exiting behaviour that defines SR2V is never exercised. Long-video benchmarks have the complementary gap, scoring scene-level coherence over an extended horizon while supplying no reference images, so subject fidelity goes unmeasured. What SR2V demands is a **condition schedule**, together with a protocol that reports quality where that schedule changes instead of averaging it over the whole video. We therefore build them as described below.

A.2.1 Evaluation Dataset

Reference images. We sample reference images from established reference-to-image/video benchmarks (Yuan et al., 2026a; Ruiz et al., 2023), preserving a category distribution as diverse as theirs: humans, animals, daily objects and products.

Prompts. To ensure that every prompt specifies an explicit entering or exiting event, we compose the prompts with Qwen3 Yang et al. (2025). For each sample we draw a set of references from the pool and ask the model to write a single narrative in which subjects enter, leave and interact with one another. The requested event follows the category of the reference: living subjects such as humans and animals enter or leave under their own motion, whereas static objects are brought into or out of the shot by camera movement, such as pulling back or panning.

Condition schedule. Once the global reference set and narrative of a sample are fixed, we split the narrative into consecutive segments, one per interaction, and pair each segment with the reference set its description requires; a segment in which a subject arrives is matched with the reference of that subject together with those of all subjects already present. Since StreamingR2V generates chunk

by chunk, this schedule is then broadcast to the chunk level, so that every evaluation sample is a sequence of per-chunk conditions, each a reference set paired with a prompt.

Statistics. The procedure yields 120 evaluation samples: 40 with 200 frames, 40 with 400 frames, 24 with 800 frames, and 16 with 1200 frames. More detailed information about entering and exiting events is shown in Tab. 5.

Table 5: **Statistics reference interaction of curated SR2V evaluation datasets.** Please be note that within one sample there may contains multiple interaction events.

Horizon	Samples	Entry Events	Exit Events
200 frames	40	34	14
400 frames	40	34	20
800 frames	24	20	14
1200 frames	16	14	9

A.2.2 Evaluation Metrics

Since an SR2V video is produced chunk by chunk, the chunk is the smallest unit that carries a condition, and chunks of the same sample may be conditioned differently. Averaging a metric over a whole video would therefore mix scores obtained under different reference sets and prompts. We instead evaluate at two levels of granularity, summarized in Fig. 8: metrics that do not depend on conditioning are computed over the entire video, whereas the R2V metrics are computed per condition and then aggregated. Building on this multi-granularity evaluation, we further propose interaction-compliance metrics that more faithfully assess reference presence at the interaction-event level, summarized in Fig. 9.

Notation. A generated sample is a chunk sequence $\mathcal{C} = (C_1, \ldots, C_N)$. The reference set and the prompt each stay constant over a stretch of chunks, but the two need not change at the same chunk, so each induces its own partition of $\mathcal{C}$. With $0 = r_0 < \cdots < r_m = N$ marking the chunks after which the reference set changes, the group $G_j^{\mathrm{ref}} = (C_{r_{j-1}+1}, \ldots, C_{r_j})$ is generated under a single reference set $\mathcal{R}_j$; the prompt boundaries p_k likewise induce groups G_k^{pmt}, each under a single prompt $\mathcal{P}_k$. We write $F(\cdot)$ for the frames sampled from a group of chunks; since all chunks are equal in length and sampled uniformly, $|F(G)|$ is proportional to the number of chunks in G.

Part 1: Global metrics. Aesthetic score, image quality, motion smoothness, background consistency and dynamic degree, all provided by VBench, describe the video as a whole and do not refer to the conditioning. We compute them over the full sequence, so that a discontinuity introduced at a chunk boundary is penalised rather than hidden by per-chunk averaging.

Part 2: Per-condition metrics. For the R2V consistency we score each group against the condition that produced it, and average the group scores over the sample weighted by the number of sampled frames, so that a condition governing a longer stretch of the video contributes proportionally more:

$$S \;=\; \frac{\sum_g |F(g)| \cdot \Psi\big(F(g),\, c_g\big)}{\sum_g |F(g)|}, \tag{7}$$

where g ranges over the groups induced by the condition that Ψ depends on: GME is evaluated on the prompt groups G_k^{pmt} against $\mathcal{P}_k$, and Face-Sim and DINO-I on the reference groups G_j^{ref} against $\mathcal{R}_j$, with Face-Sim applied to human references and DINO-I to the remaining categories. Pairing G_j^{ref} with $\mathcal{R}_j$ rather than with the union of all references of the sample is what keeps the measurement honest: a subject is only required to be present while its reference is actually part of the condition.

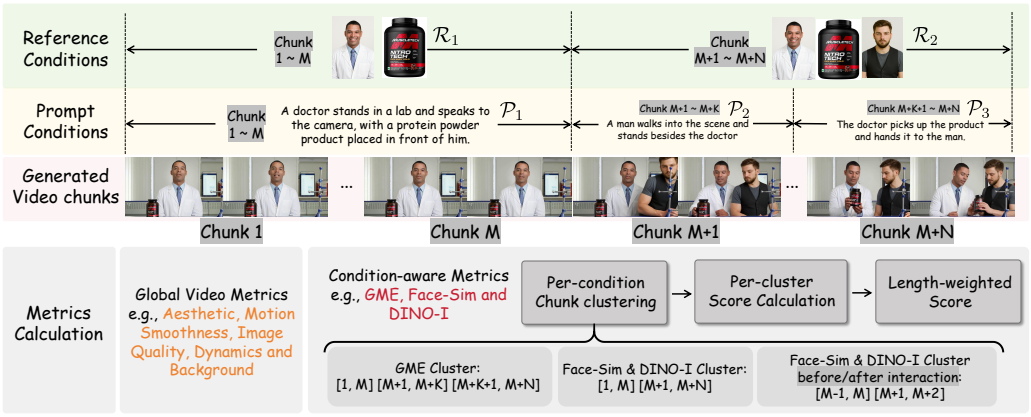

Figure 8: **Overview of the SR2V evaluation consist of different levels of granularity.** An evaluation sample carries two condition tracks that run along the same timeline but need not change together: the reference set (*first row*) and the prompt (*second row*). The example shown spans $M{+}N$ chunks in total, the quantity denoted N under *Notation*. Its reference set changes once, giving the groups $[1, M]$ and $[M{+}1, M{+}N]$ under $\mathcal{R}_1$ and $\mathcal{R}_2$, while the prompt changes twice, giving $[1, M]$, $[M{+}1, M{+}K]$ and $[M{+}K{+}1, M{+}N]$ under $\mathcal{P}_1$ to $\mathcal{P}_3$; the second prompt boundary at $M{+}K$ falls inside the second reference group, so the two tracks cut the generated chunk sequence (*third row*) into two different groupings. Metrics (*fourth row*) are computed at matching granularities: conditioning-free video quality over the whole sequence, GME per prompt group and Face-Sim and DINO-I per reference group, each group scored against the condition that produced it and aggregated by sampled-frame count. At the reference change M we additionally score the two chunks before and after it, (C_{M-1}, C_M) against $\mathcal{R}_1$ and (C_{M+1}, C_{M+2}) against $\mathcal{R}_2$, which isolates the identity fidelity of the newly arrived reference.

Multi-subject matching. The reference implementations of Face-Sim and DINO-I in OpenS2V-Eval keep only the largest detection in a frame and compare it against each reference independently, which is undefined once several subjects share a frame and allows two references to claim the same subject. We therefore keep every detection and resolve the correspondence explicitly. Let $\sigma(\rho, d)$ be the embedding cosine similarity between a reference $\rho \in \mathcal{R}_j$ and a subject d detected in a sampled frame f. We match them by $\pi_f = \arg\max_\pi \sum_{(\rho,d)\in\pi} \sigma(\rho, d)$, where π ranges over the matchings that pair each reference with at most one detection and each detection with at most one reference, so that two references can never collapse onto the same subject; the maximum is attained exactly by the Hungarian algorithm. Writing F_ρ for the frames of the group in which ρ is matched, the subject score averages per reference first and then across references,

$$\Phi\big(F(G_j^{\text{ref}}),\, \mathcal{R}_j\big) \;=\; \frac{1}{|\mathcal{R}_j|} \sum_{\rho\in\mathcal{R}_j} \frac{1}{|F_\rho|} \sum_{f\in F_\rho} \sigma\big(\rho, \pi_f(\rho)\big), \tag{8}$$

so that a reference visible in only a few frames is not down-weighted. A reference that is never detected contributes 0 rather than being dropped from the outer mean, since discarding it would reward a model for omitting the subject altogether.

Interaction consistency metrics. Eq. 7 still averages over a whole sample, which reveals little about the moment that defines SR2V, since a brief failure at a reference change is diluted by the chunks around it. We therefore report a second set of scores on a window of four chunks centred on each boundary r_j, split into the two chunks before it, $B_j = (C_{r_j-1}, C_{r_j})$, and the two after it, $A_j = (C_{r_j+1}, C_{r_j+2})$. The two halves are deliberately scored against different conditions, Φ_{before} averaging $\Phi(F(B_j), \mathcal{R}_j)$ over the boundaries of the benchmark and Φ_{after} averaging $\Phi(F(A_j), \mathcal{R}_{j+1})$. Since $\mathcal{R}_{j+1}$ is the set that contains the newly arrived reference, the gap $\Phi_{\text{before}} - \Phi_{\text{after}}$ measures how much identity fidelity the transition costs. Background consistency and motion smoothness are instead computed over $B_j \cup A_j$ as a whole rather than on either half, as a discontinuity introduced at r_j is only visible across the boundary. These are the scores reported in Tab. 2 and Tab. 3 of the ablation study.

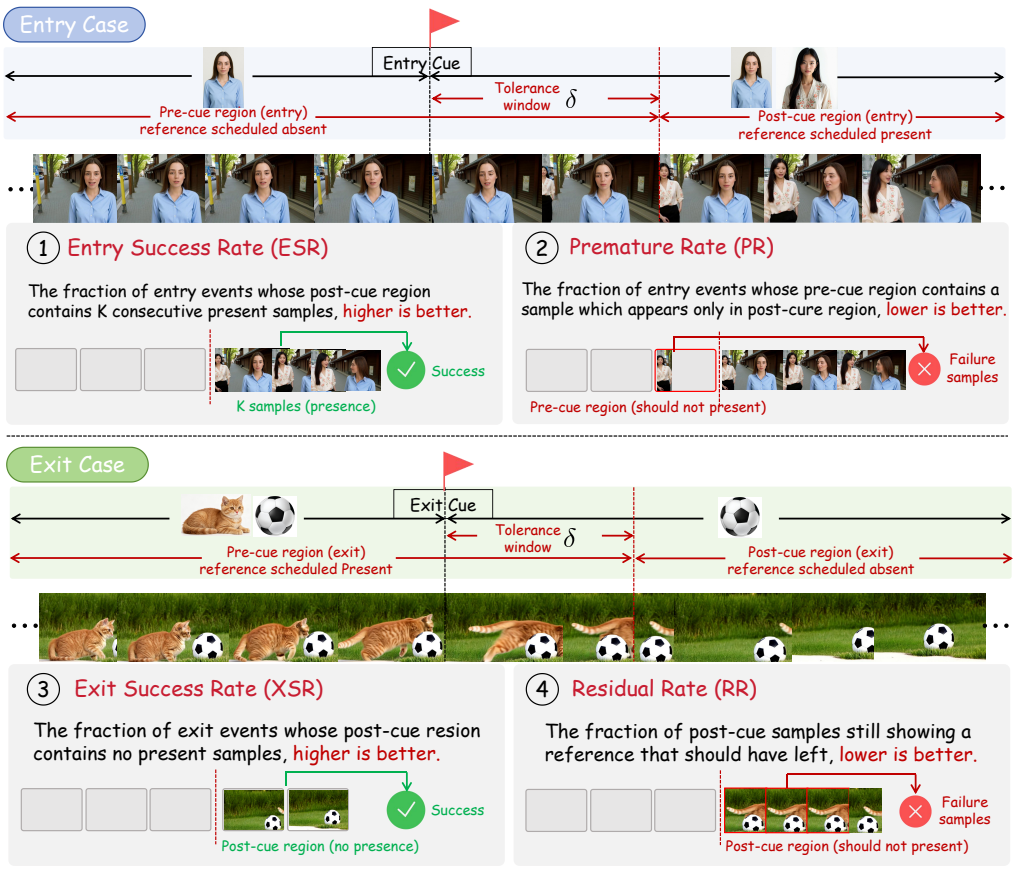

Figure 9: **The evaluation pipeline for the metrics of interactive compliance.** We compute the entry success rate (ESR) and premature rate (PR) to evaluate whether, during an interaction that adds a reference (an entry event), the corresponding subject appears only after the cue and does not appear prematurely. When removing a reference (an exiting event), we use the exiting success rate (XSR) and residual rate (RR) to evaluate whether the subject disappears after the cue and no longer remains present. A tolerance window δ after the cue is introduced to provide the model with a margin for the subject to fully enter or exit the scene.

Part 3: Interaction Compliance. Eq. 7 and Eq. 8 score each reference only within the group where it is included in the condition. However, a subject rendered long before its reference is introduced is not penalized, since those frames belong to groups whose conditions exclude that reference. Similarly, a subject that remains in the frame after its reference is withdrawn also escapes penalty. As a result, these cases may still receive high consistency scores while failing to follow the scheduled reference set, which determines whether a reference should appear or be absent before and after each interaction. Addressing this issue is crucial for evaluating SR2V, and we therefore plan to introduce an interaction-event-based scoring criterion.

Events, Presence, and Tolerance. A reference added at a chunk boundary is an *entry event*, one that is removed is an *exit event*, and the boundary frame is the event *cue*. A reference counts as *present* in a sampled frame when the detection matched to it reaches a similarity threshold of τ. One departure from Eq. 8 is essential: matching is performed against the *union* of the sample's references rather than the set in force at that frame, because premature appearances and post-exit residues are only observable when a reference is deliberately absent from the condition. Since the entry and exit of a reference naturally take some time, we ignore a *tolerance window* after the cue, spanning a fraction δ of the chunk it opens. Entry and exit events use the same tolerance window, which keeps their rates comparable. This leaves the two regions shaded in Fig. 9: a *pre-cue region*, where the

reference is scheduled to be absent, and a *post-cue region*, where an entry is scheduled to be present and an exit is scheduled to be absent. Finally, since a single-frame decision is susceptible to detector flicker, a reference is counted as appearing only over a run of K consecutive present samples. We set $\delta = 0.25$, $K = 3$, and $\tau = 0.30$ for each backend.

The four rates for interaction compliance. As Fig. 9 shows, we report the ***entry success rate***, the fraction of entry events whose post-cue region contains such a run; the ***premature rate***, the fraction of entry events whose pre-cue region contains such a run; the ***exit success rate***, the fraction of exit events whose post-cue region contains no such run; and the ***residual rate***, the mean fraction of post-cue samples that still show a reference that should have left. The entry and exit criteria are mirror images: one requires a sustained run, while the other requires its absence, which keeps the rates commensurable. The premature and residual rates are not restatements of the entry or exit success rates; they are calculated from sampled frames and therefore reflect how early a reference appears before its entry cue, as well as how long it remains visible after its exit cue.

A.2.3 Baseline Setup

To the best of our knowledge, few published methods explicitly control subject entry and exit by changing the reference set during a long rollout, making direct comparison with our system challenging. We therefore set up 3 families of baselines, each covering the horizon at which a comparison is meaningful.

Commercial models. Customized generation at the horizon of 200 frames is within reach of popular commercial systems (MiniMax, 2026; Kling, 2026), so the comparison at this length can directly include them. Their interfaces, however, accept the conditioning only once, before generation begins, and a schedule cannot be expressed through them. We therefore supply the prompt as a single narrative, before it is split into segments, together with the union of every reference that appears anywhere in the sample. We stress that this follows from the API rather than from a restriction we impose, and that it has a visible consequence: a subject scheduled to arrive later is available from the first frame and tends to be rendered there, as Fig. 4 shows.

Two-stage baselines. Beyond 200-frame video generation we approximate SR2V by composing two models that each solve half of the task: a short-clip R2V model, which renders a single chunk from a reference set, and SVI, which is capable of extending a video chunk by chunk and takes a separate prompt for each. Generation follows the reference schedule. At every boundary r_j we invoke the R2V model with the condition of that chunk, and from there until the next boundary we let SVI continue the video it produced; a prompt change falling inside a reference group is handled by SVI alone. The pipeline therefore has a basic form of interactive control. We pair VACE (Jiang et al., 2025), Phantom (Liu et al., 2025), MAGREF Deng et al. (2026), and RefAlign (Wang et al., 2026a) with the original SVI, each built on Wan2.1-14B, which matches the backbone of StreamingR2V so that the comparison reflects the method rather than the capacity of the foundation model. Re-invoking the R2V model at a reference change resets the rollout: that chunk is generated from the references alone, free of accumulated drift, whereas StreamingR2V renders it continuously. The baselines are thus favoured at precisely the chunks Tab. 2 isolates, making those margins conservative. The reset instead costs them continuity with the preceding scene.

SVI + R2V SFT. The aforementioned commercial model can generate only relatively short clips, while the two-step baselines are not jointly trained with history and reference guidance. To provide a more comprehensive comparison and better justify the effectiveness of our methodology, we implement a naive baseline that follows the input paradigm described in Sec. 3.2.1 and uses a history error bank similar to the video latent error bank in SVI. This baseline is finetuned on the same dataset as StreamingR2V and ablates the key designs of StreamingR2V, including the reference-reflective value gate and the virtual rollout error chain. However, since it accepts joint history and reference conditions and is trained on the same dataset as StreamingR2V, it is expected to support the SR2V task and is therefore directly comparable to StreamingR2V.

A.3 Extended Experiments

A.3.1 Additional Qualitative Comparisons

In this section, we provide additional visualizations of qualitative comparisons. Fig. 10 shows two cases of 200-frame SR2V generation results, while Figs. 11 and 12 show two examples of 800-frame SR2V generation results. Although StreamingR2V consistently achieves better overall consistency, including subject fidelity and background consistency, its advantages become particularly noticeable under more challenging interaction schedules. For instance, when both entering and exiting interactions occur in the same sample, as shown at the bottom of Fig. 10, StreamingR2V faithfully renders the proper behavior of the target subject in accordance with the user prompt. In contrast, commercial models that accept all conditions only up front cannot faithfully render entering and exiting behaviors according to the interaction schedule. They tend to include all subjects from the beginning of the video and exhibit a relative delay when subjects exit. In Figs. 11 and 12, the reference interactions are more complicated, and the relationships among subjects are also more complex, e.g., *a man enters the frame and a dog jumps into the man's arms*. The two-stage baseline still suffers from abrupt background changes and sudden subject appearances, while the SVI+R2V SFT baseline suffers from low subject fidelity after reference interaction. In contrast, StreamingR2V not only demonstrates high fidelity to the interaction schedules, but also produces better multi-reference interaction patterns that follow the prompt. Overall, these visualization results further demonstrate the robustness of StreamingR2V under more challenging conditions.

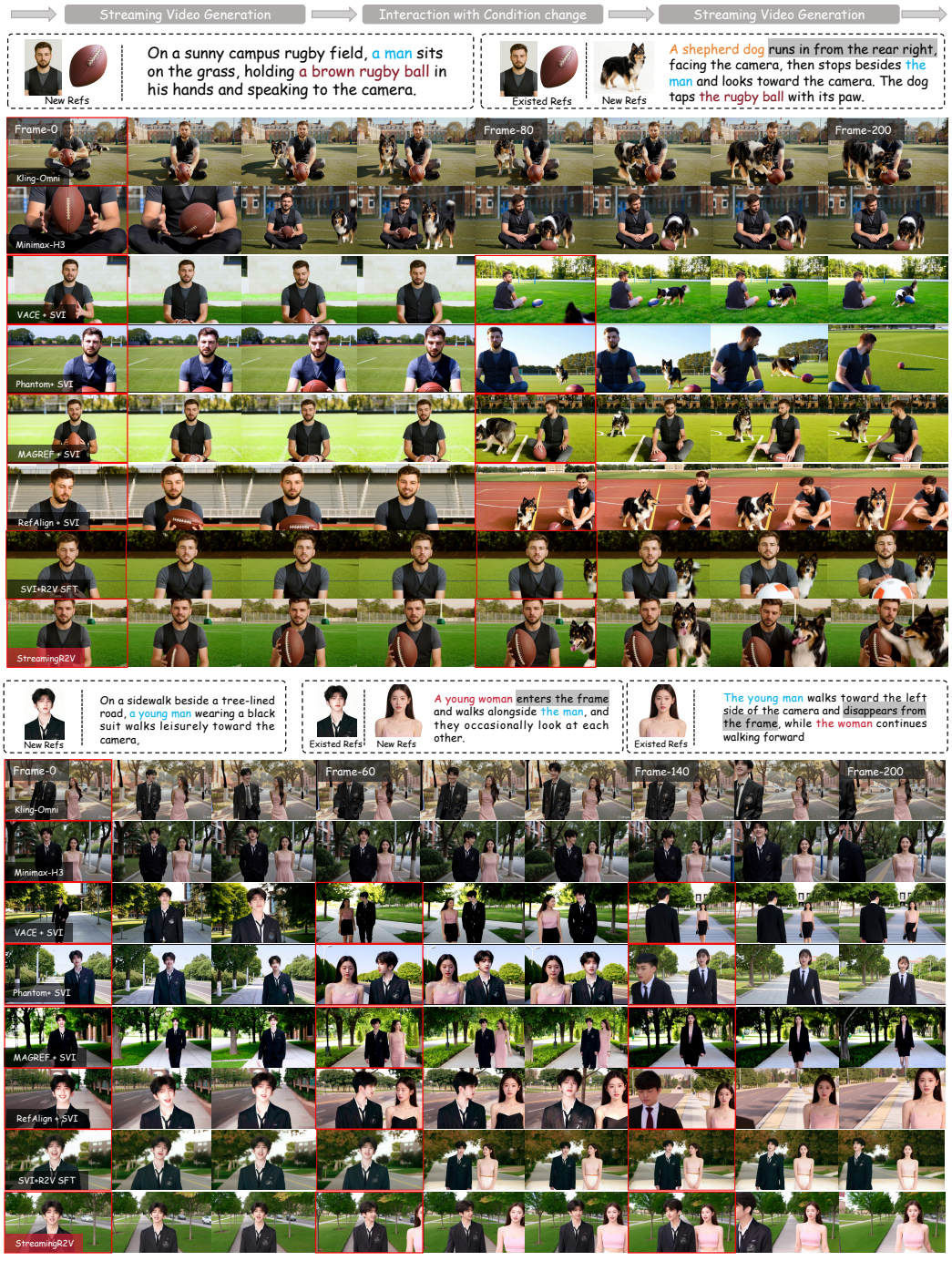

Figure 10: **More qualitative comparison results on the horizon of 200 frames.** Diverse reference subjects and complex entering and exiting interactions are introduced, StreamingR2V shows advantages on background consistency and ideal prompt following, especially when multiple subjects are included and have specific interaction patterns. The initial frames for each interactive condition setting are highlighted with the red box. (*Zoom in for better view*).



Figure 11: **More qualitative comparison results on the horizon of 800 frames.** Diverse reference subjects and complex entering and exiting interactions are introduced, StreamingR2V shows advantages on background consistency and ideal prompt following, especially when multiple subjects are included and have specific interaction patterns. The initial frames for each interactive condition setting are highlighted with the red box. (*Zoom in for better view*).

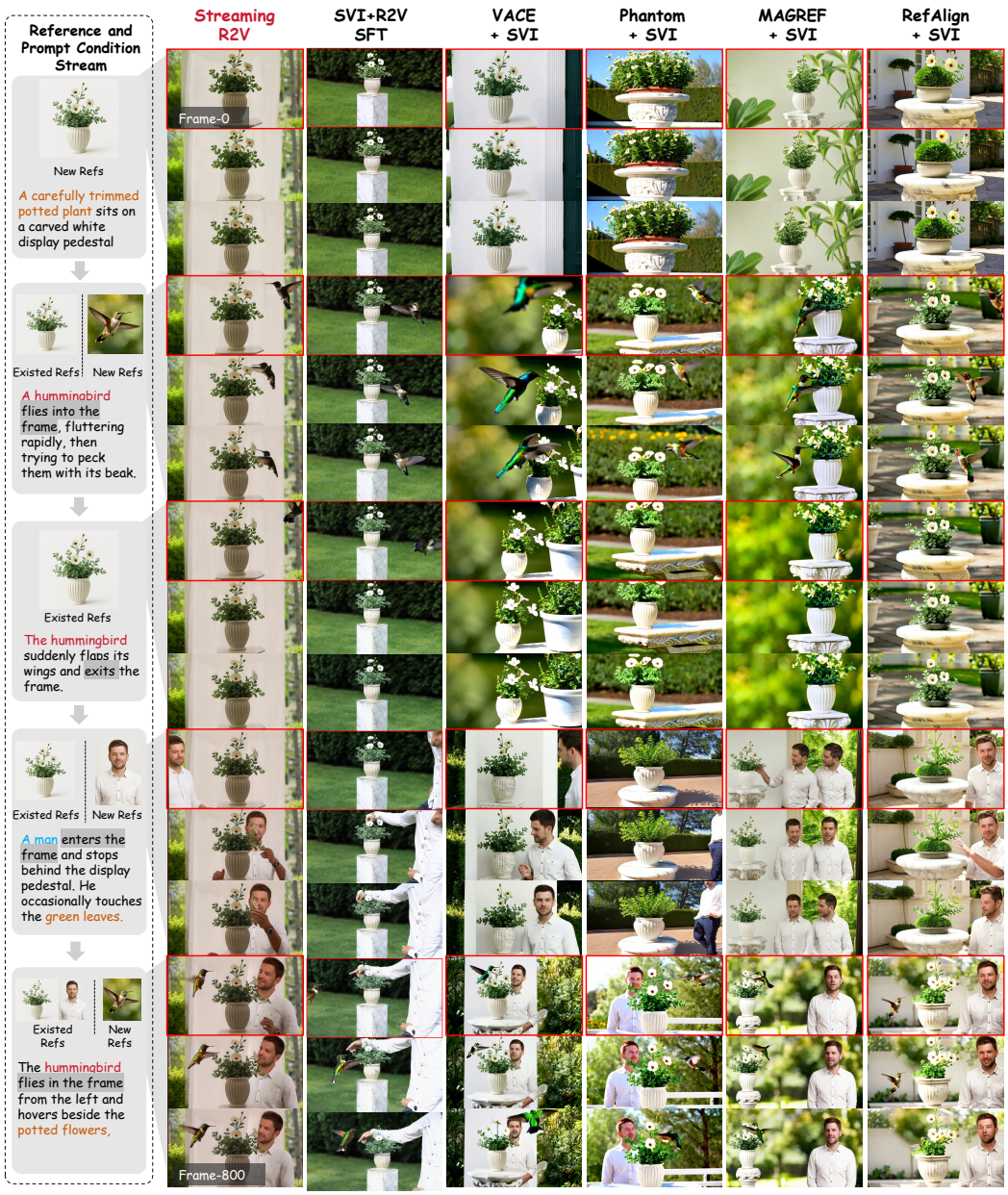

Figure 12: **More qualitative comparison results on the horizon of 800 frames** Diverse reference subjects and complex entering and exiting interactions are introduced, StreamingR2V shows advantages on background consistency and ideal prompt following, especially when multiple subjects are included and have specific interaction patterns. The initial frames for each interactive condition setting are highlighted with the red box. (*Zoom in for better view*).

A.3.2 Additional Ablation Studies

We provide additional ablation studies on key hyperparameter choices in StreamingR2V's core design, including the projection dimension d_{id} of the reference reflective value gate (hereafter referred to simply as the gate) and the cap depth ℓ_{max} of the virtual rollout error chain.

ID Subspace dimension d_{id} of reference reflective value gate. Tab. 6 and Fig. 13 show the quantitative and qualitative comparisons of different choices of d_{id}, respectively. Both results show that a smaller choice of d_{id} still outperforms the setting without this design. However, our current choice of $d_{id} = 64$ achieves the best results among the three settings, especially in preserving the fidelity of the reference subject when it enters the frame after the interaction cue.

Table 6: Additional ablation study results for d_{id} in the reference reflective value gate. All settings are evaluated with the original virtual rollout error chain design, where $\ell_{max} = 3$.

d_{id}	MS↑	BC↑	FS_b↑	FS_a ↑	$DINO_b$ ↑	$DINO_a$ ↑
w/o gate	<u>0.992</u>	0.916	0.528	0.414	0.680	0.515
32	0.991	<u>0.918</u>	<u>0.578</u>	<u>0.513</u>	<u>0.699</u>	<u>0.611</u>
64 (Ours)	**0.993**	**0.925**	**0.601**	**0.533**	**0.713**	**0.632**

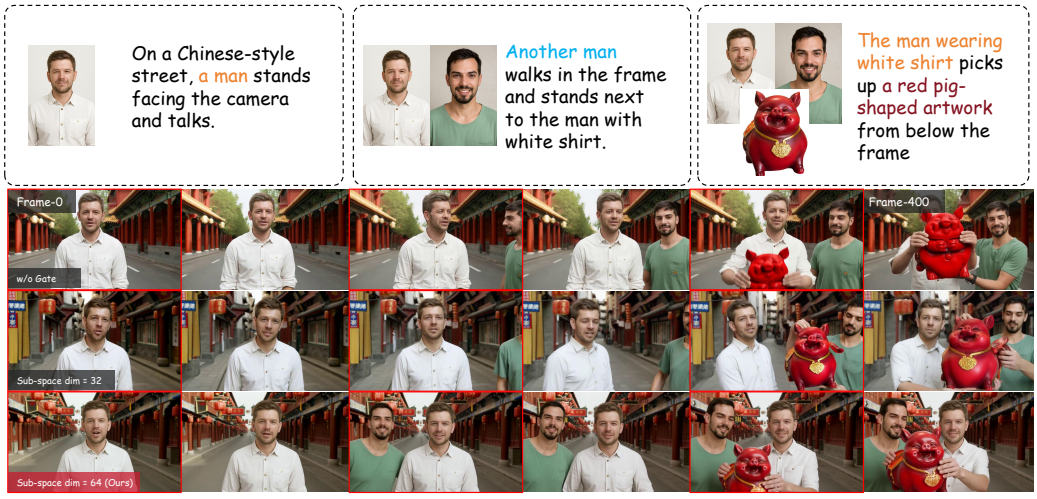

Figure 13: **Qualitative comparison results on different choice of d_{id}** The initial frames for each interactive condition setting are highlighted with the red box.

Cap depth ℓ_{max} of virtual rollout error chain . Tab. 7 and Fig. 14 present quantitative and qualitative comparisons for different choices of ℓ_{max}, ranging from 1 to 4, where $\ell_{max} = 1$ denotes the original error bank used in SVI. The results demonstrate that introducing a virtual rollout error chain consistently alleviates the inconsistency between background coherence and subject fidelity, particularly chunks involving reference interaction. When $\ell_{max} < 3$, noticeable background shifts still occur, and the incomplete preservation of reference cues caused by these shifts also leads to lower subject consistency. Although the setting $\ell_{max} = 4$ further improves motion smoothness and background consistency, the gain is only marginal. Therefore, we consider that, under our training setting, $\ell_{max} = 3$ is sufficient for StreamingR2V to achieve near-ceiling performance.

A.3.3 Runtime Analysis

Given the hidden-state dimension of Wan-14B, which is 5120, and considering that W_{id} accounts for the majority of the trainable parameters in the reference-reflective value gate, the total number of additional parameters introduced by the value gate is

$$5120 \times 64 \times 40 \approx 13.1\,\mathrm{M}.$$

Table 7: Additional ablation results for the cap depth ℓ_{max} of the virtual rollout error chain. All settings are evaluated with the original reference reflective value gate, where $d_{id} = 64$.

ℓ_{max}	MS ↑	BC ↑	FS_b ↑	FS_a ↑	$DINO_b$ ↑	$DINO_a$ ↑
1 (Original SVI)	0.993	0.913	0.555	0.493	0.681	0.614
2	0.992	0.914	0.570	0.481	0.692	0.576
4	**0.994**	**0.929**	0.590	**0.539**	0.697	0.614
3 (Ours)	0.993	0.925	**0.601**	0.533	**0.713**	**0.632**



Figure 14: **Qualitative comparison results on different choice of** ℓ_{max} The initial frames for each interactive condition setting are highlighted with the red box.

Since the reference-reflective value gate is a newly introduced module, we further analyze its inference cost under the setting described in Tab. 6. All inference time measurements are conducted for generating a single 81-frame chunk at a resolution of 832×480 with 50 denoising steps on a single NVIDIA A100 GPU. In this analysis, the number of references is fixed at 2, and each setting is evaluated five times, with the average inference time reported. As shown in Tab. 8, introducing the reference-reflective value gate does not lead to any noticeable additional computational overhead.

Table 8: Inference runtime analysis of reference reflective value gate under different setting of d_{id}

d_{id}	Additional Parameter (M)	VRAM (GB)	Time (second)
w/o gate	0	39.51	691.34
32	≈7	39.57	698.95
64	≈13	39.65	706.52

A.3.4 HUMAN EVALUATION PREFERENCE

We invite 30 volunteers to compare **StreamingR2V** with four baselines. Each volunteer evaluates 16 cases randomly sampled from the 40 200-frame evaluation dataset, with 5 method outputs per case (80 videos in total). The guideline is shown in Tab. 9:

Table 9: Guidelines of human preference evaluation.

Part 1 Task Background This study focuses on reference-image-conditioned long-video generation, where the video can last for dozens of seconds. Given several reference images of target subjects and a text instruction, the model generates a long video and supports dynamically adding or removing reference subjects during generation, allowing them to enter or leave the scene at the specified moments in the instruction. This task closely reflects real-world creative workflows such as advertising, short dramas, and multi-character storytelling. Each case below presents outputs from different methods under the same conditions. Please provide your subjective preference evaluation.
Part 2 Survey Guidelines Each case includes reference images, a text instruction, and 5 anonymous samples. Please select the best sample for each of the following three dimensions, for a total of 3 votes. These three dimensions should be judged independently, and the same sample may be selected more than once. **1. Overall Fluency**: The background and scene remain consistent over time, or follow the camera movement fluently. The subjects' motions are natural and coherent, there are no camera cuts, frame jumps, or abrupt lighting changes, and there is no obvious degradation in visual quality over long durations. **2. Instruction Following**: The visual content and actions match the text instruction. Pay special attention to whether subjects enter or leave the frame naturally, and whether they appear too early or fail to appear in time. **3. Subject Fidelity**: The similarity between the subjects and the reference images. For human subjects, focus on facial features, hairstyle, and clothing; for other objects, focus on overall appearance, shape, color, and texture. Also note whether this similarity remains stable throughout the entire video, before and after new subjects enter.
Please do not select a sample for other dimensions simply because it is better in one dimension. If the differences are hard to distinguish, choose the relatively better one rather than leaving it blank.

The human-based preference statistics are shown in Fig. 15. It shows that StreamingR2V shows advantages over the competitive commercial models such as MiniMax-H3 in three metrics, and significantly over several engineering-based two-step baseline. This results further consolidate the advantage of StreamingR2V over these baselines.

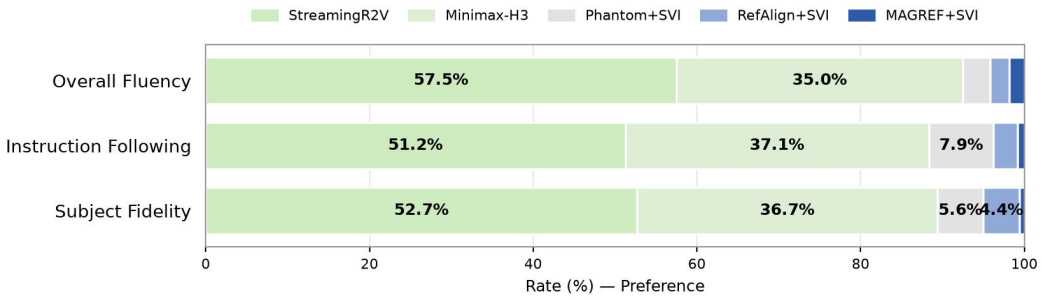

Figure 15: **Human evaluation preference results.** StreamingR2V outputperforms all other baselines including competitive commercial models such as MiniMax-H3.

A.4 More Results

We provide additional visualizations of raw videos on our anonymous demo page, `https://sr2v-demo.github.io/`, and encourage readers to visit it. Please also note that the original video samples for all qualitative comparison results discussed in both the main submission and the Appendix are available on the demo webpage.

A.5 Discussions

A.5.1 Practical Application

Customized Long-Take Generation. StreamingR2V is capable of generating long, stable videos with high subject fidelity. This makes it particularly promising for applications that require extended single-shot video generation, such as livestreaming scenarios where a user remains in the frame while interacting with different products or meeting different people. StreamingR2V is therefore well suited to such highly customized generation scenarios.

World Generation with Customized Visual Elements. StreamingR2V also demonstrates strong performance in long-video generation when camera movement is evolved, where different visual elements can be selectively included in or excluded from the scene. This capability closely resembles the *exploration* property of world models. As a result, StreamingR2V has strong potential for applications in customized world generation, data synthesis, and related tasks.

A.5.2 Limitations and Future Work

Built upon Wan-I2V-14B, StreamingR2V inevitably inherits the limitations of its foundation model. Moreover, since StreamingR2V primarily focuses on the visual performance of SR2V, such as interaction compliance and subject fidelity, this work does not extensively explore real-time interaction techniques, such as few-step distillation. In the future, we plan to incorporate such techniques to make the system more practical for real-world applications.