

ReferStream: Interactive Subject Injection for Streaming Video Generation

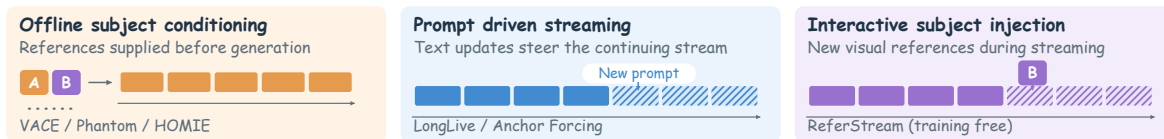
Shuolin Xu^{1,5}, Guanyuan Li^{2,5}, Boyu Li³, Peng-Tao Jiang⁴
 Hongchuan Yu¹, Xiaobin Hu⁵, Shuicheng Yan⁵

¹Bournemouth University, NCCA UK ²Zhejiang University
³HKUST ⁴vivo BlueImage ⁵National University of Singapore

Abstract

Injecting new subjects naturally during streaming video generation requires integrating newly supplied visual references into the existing video while preserving established subjects and the scene. Encoding references independently of the current scene provides little guidance for this integration, while keeping their features in persistent memory can interfere with established context. We propose ReferStream, a framework that requires no additional training and combines reference construction using the current scene with transient cache injection. Joint Latent places reference latents in target regions of the latest scene latent, allowing reference features to be extracted in context. Cache Restoration uses these features to condition the injection chunk, then restores the overwritten global cache entries while retaining the newly generated history. This procedure incorporates new visual conditions without retraining the generator or revising emitted frames. We further introduce ReferStream Bench, a new benchmark with scheduled reference arrivals, evolving instructions, and explicit annotations of subject presence for InShot and MultiShot generation. Evaluation follows the event schedule and measures both the fidelity of new subjects and the preservation of established subjects, alongside video quality, text adherence, and count accuracy. Experiments on ReferStream Bench show higher scores for text adherence, subject similarity, and count accuracy than the compared baselines, including offline methods given all references in advance.

When can a new visual subject enter the generation process?



ReferStream: interactive subject injection on a frozen generator

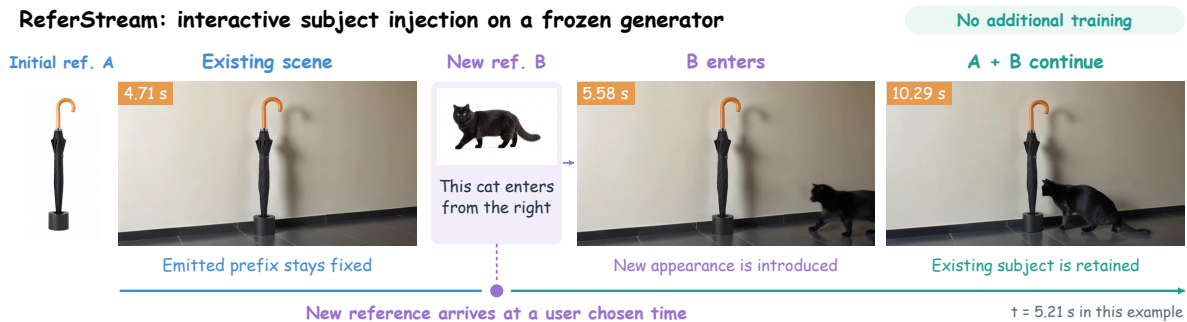


Figure 1 Interactive subject injection during streaming generation. A new cat reference is supplied after the umbrella scene is established. The cat enters while the umbrella remains and previously emitted frames stay fixed.

1 Introduction

Injecting new subjects into a video during streaming generation requires integrating newly supplied visual references into a scene that has already been established. A user may provide a photograph of a cat after watching a lobby scene containing an umbrella, then ask the cat to enter from the right (Figure 1). The reference specifies the cat’s appearance, but its introduction must also fit the ongoing scene and retain the umbrella. With further interactions, multiple references may arrive at different times. Each update must guide subsequent generation while leaving the video the user has already seen unchanged.

Autoregressive video diffusion provides a basis for this interaction by generating successive chunks conditioned on cached history [11, 29]. LongLive and LongLive2 support changing text instructions during generation [5, 26], and Anchor Forcing improves memory continuity across prompt changes [27]. Visual reference methods such as Phantom, VACE, and HOMIE provide control over subject appearance [1, 14, 19]; AlcheMinT further specifies temporal intervals for reference subjects [8]. We study interactive subject injection, where new visual references become available during causal generation. The arrival of a reference determines when its appearance information becomes accessible, in addition to when the subject should appear. This setting requires adapting an existing scene to information supplied after generation has begun.

Two difficulties arise when incorporating these references into a streaming generator. First, a reference depicts its subject at a scale, viewpoint, and background that may differ from the generated scene. Encoding it independently supplies appearance information without explicitly relating it to the surrounding content. Second, the generator’s cache already represents established subjects and the scene. Replacing persistent entries with reference features can help introduce a subject, but leaving those features in place may continue to favor the reference composition after the subject has entered. Reference construction and the duration of its influence therefore need to be considered together.

We propose ReferStream, which combines Joint Latent and Cache Restoration on a frozen LongLive2 backbone [5]. Joint Latent places encoded references within target regions of the latest clean scene latent, allowing reference features to be extracted in context. Cache Restoration temporarily writes these features into selected global cache entries, then restores the original entries after the injection chunk is generated and committed. The newly generated history remains available to carry the subject forward, without updating model parameters or revising emitted frames.

Assessing this task requires accounting for reference arrival and the persistence of established subjects. OpenS2V-Nexus evaluates reference consistency, naturalness, and text relevance [30]. We introduce ReferStream Bench with scheduled reference arrivals, evolving instructions, and expected subject presence. InShot introduces subjects within a continuous shot, while MultiShot permits declared transitions. Evaluation measures new and established subjects alongside video quality, text adherence, and count accuracy.

Our contributions are as follows:

- We formulate interactive subject injection during causal video generation, where visual references arrive at user chosen times and must be integrated while preserving established subjects and scene content.
- We develop ReferStream, combining Joint Latent for reference construction in scene context with Cache Restoration for transient conditioning, without additional training or revision of emitted frames.
- We introduce ReferStream Bench for InShot and MultiShot generation, with explicit event schedules and evaluation of new subject fidelity, established subject preservation, video quality, text adherence, and count accuracy.
- Experiments show higher text adherence, subject similarity, and count accuracy than the compared baselines, including offline methods given all references in advance. Video quality results are mixed. Ablations characterize the tradeoffs between fidelity and preservation, and experiments on OpenS2V [30] further support Joint Latent.

2 Related Work

Video generation models. Spatiotemporal latent compression and diffusion transformers underpin recent video generators such as CogVideoX, HunyuanVideo, and Wan [16, 25, 28]. LTX-Video emphasizes efficient latent synthesis [9], while LTX-2 and MiniMax H3 extend generation to synchronized audiovisual content and multimodal conditioning [10, 21]. These advances provide strong synthesis backbones. Our work concerns how a causal generator incorporates visual references that become available after part of the video has already been emitted.

Subject-to-video generation. Subject-to-video generation uses visual references to specify identities beyond text descriptions. Video Alchemist and ConceptMaster study personalization across arbitrary subject categories and the separation of multiple reference concepts [4, 12]. Phantom and MAGREF develop reference conditioning mechanisms for consistent synthesis of one or more subjects [6, 19]; VACE unifies reference generation and video editing, while HOMIE focuses on personalization of humans and objects [1, 14]. OpenS2V-Nexus provides a detailed benchmark and a large dataset for evaluating and training models conditioned on subject references [30]. AlcheMinT further conditions references on temporal intervals to control subject appearance and disappearance [8]. Our setting instead reveals new references during causal generation, requiring their integration into an established scene while preserving subjects that should remain. Reference arrival therefore determines what information the generator can use, in addition to when a subject should appear.

Streaming generation and interactive editing. Diffusion Forcing and SkyReels-V2 use heterogeneous temporal noise levels for sequence generation and long video synthesis [2, 3]. CausVid distills a causal video generator, while Self Forcing trains on context generated by the model itself to reduce exposure bias [11, 29]. FramePack compresses historical context for efficient continuation [31]. CausVid also supports streaming video-to-video translation; StreamDiffusion and StreamDiffusionV2 optimize interactive pipelines for low latency [7, 15, 29]. LongLive supports prompt changes through cache updates, and LongLive2 extends interactive generation to videos with multiple shots with efficient parallel training and inference [5, 26]. Anchor Forcing improves continuity across prompt changes by using anchors to guide recaching and realigning positional embeddings [27]. Building on this streaming formulation, we study how newly supplied visual references enter the cache and how the original global memory is restored after injection, without revising emitted frames.

3 Method

3.1 Autoregressive Video Diffusion and Baseline

Autoregressive video diffusion denoises successive latent chunks while keeping the generated prefix clean, allowing different noise levels across time [2]. We use chunk index k and diffusion time u , with $u = 0$ denoting clean latents. The frozen backbone is LongLive2-5B [5]. Given prompt p_k , cut flag $c_k \in \{0, 1\}$, and committed cache $\mathcal{M}_{k-1}$, its update is

$$\begin{aligned} (\hat{z}_k, \mathcal{M}_k^{\text{den}}) &= \mathcal{D}_\theta(\epsilon_k, p_k; \mathcal{M}_{k-1}, c_k), \\ \mathcal{M}_k &= \mathcal{U}_\theta(\mathcal{M}_k^{\text{den}}, \hat{z}_k, p_k; c_k). \end{aligned} \tag{1}$$

Here $\epsilon_k \sim \mathcal{N}(0, I)$ and $\mathcal{D}_\theta$ returns both the sampled chunk and the cache after denoising. The latter includes temporary writes from the current chunk. $\mathcal{U}_\theta$ commits a clean forward pass and performs cache maintenance. History combines a global sink G , a shot sink S , and a rolling window L , with overlapping tokens deduplicated [5, 26]. We modify reference conditioning and selected global entries while keeping model parameters fixed.

3.2 Interactive subject injection

Users may submit visual references at arbitrary times, with each update applied at the next chunk boundary. We index boundary updates by j and write $e_j = (k_j, \mathcal{R}_j, \mathcal{B}_j, p_j^e, c_j^e)$. Here k_j is the boundary, $\mathcal{R}_j = (r_{j1}, \dots, r_{jQ_j})$ is an ordered reference list, $\mathcal{B}_j$ contains the corresponding normalized target boxes, and p_j^e, c_j^e are the event instruction and cut flag. Boxes are supplied explicitly or resolved from layout defaults. Each boundary has one update grouping the references assigned to it, which may include previously

Injecting visual references during streaming video generation

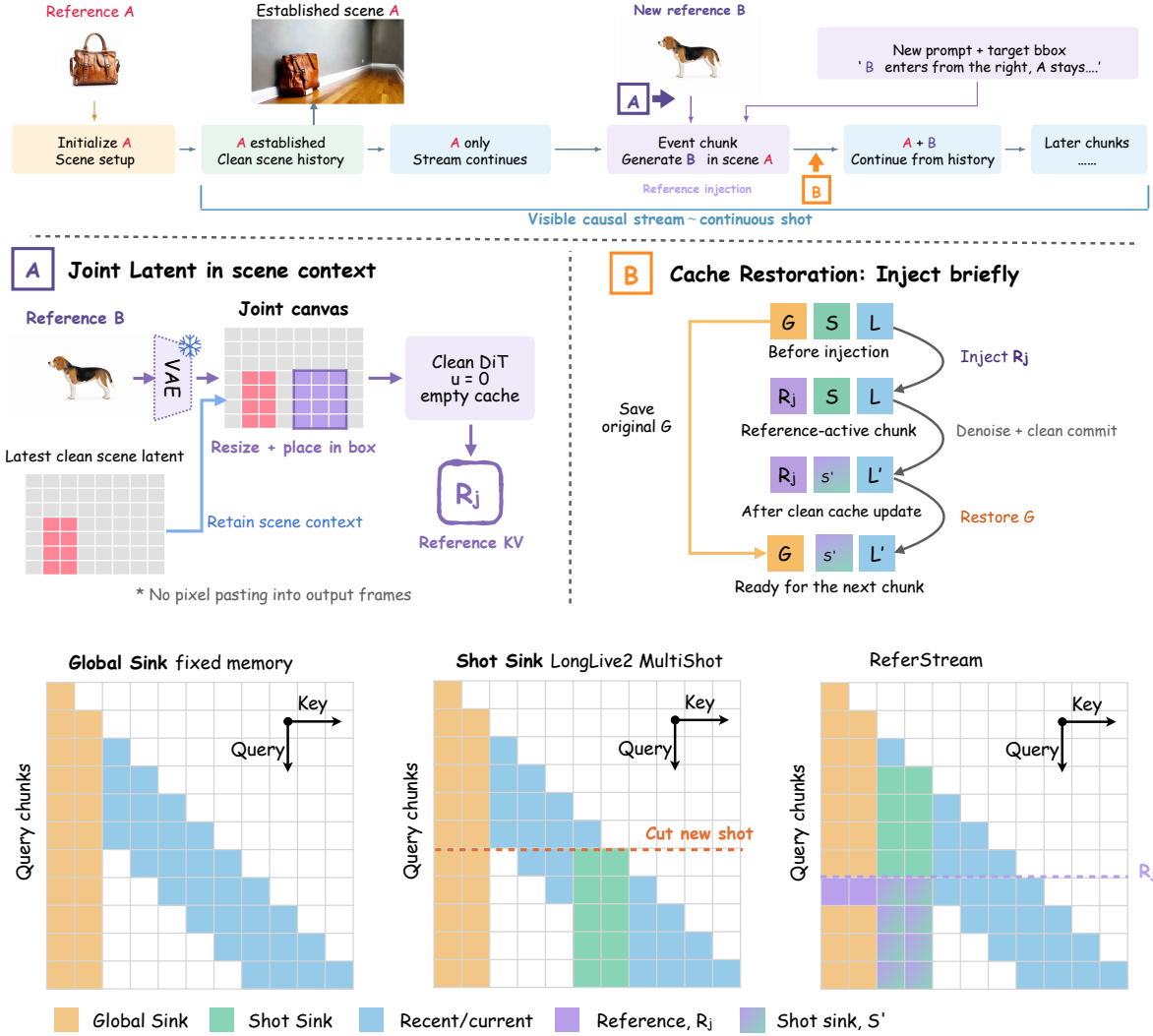


Figure 2 Overview of ReferStream, illustrated for an InShot event. Top: reference B arrives after subject A is established and conditions the next chunk. Middle: (A) Joint Latent composes the reference with the latest scene latent; a clean DiT pass extracts R_j . (B) Cache Restoration saves global K/V, injects R_j , denoises and commits the event chunk, then restores the saved entries. Generated local history L' remains. S' denotes the shot sink after the update; $S' = S$ when no refresh is scheduled for this chunk. Bottom: schematic cache access, not measured attention weights. The photographs illustrate a benchmark example; subsequent generation is schematic.

revealed images. At the event boundary, $p_{k_j} = p_j^e$ and $c_{k_j} = c_j^e$. Between updates the prompt persists and $c_k = 0$. Updates containing only text have $Q_j = 0$ and skip reference injection.

At chunk k , the available history is $\mathcal{E}_{\leq k} = \{e_j : k_j \leq k\}$. Reference conditioning depends only on this history and the generated prefix; future references are unavailable and emitted frames remain fixed. InShot uses $c_j^e = 0$, so a prompt update alone does not cut the shot. MultiShot permits $c_j^e = 1$ and the associated shot cache update. Expected subject presence defines evaluation targets, not additional identity features. LongLive2 receives the same event captions and cut flags without visual references.

An isolated reference supplies appearance but may differ from the current scene in scale, viewpoint, and background. We therefore contextualize it before injection and limit its direct influence in persistent memory. Figure 2 separates these operations into Joint Latent and Cache Restoration.

3.3 Joint Latent in scene context

Batch dimensions are omitted below. Let $h_j \in \mathbb{R}^{C \times H \times W}$ be the latest clean latent frame in the generated prefix, and let $E(r_{jq})$ be a reference image’s frozen VAE encoding. Repeating h_j across m latent frames gives a canvas $H_j \in \mathbb{R}^{m \times C \times H \times W}$; when no history exists, $H_j = 0$. Each reference encoding is repeated temporally, resized to its target rectangle, and embedded in a tensor P_{jq} of the same shape, with zeros outside that rectangle. Its box mask is $M_{jq} \in \{0, 1\}^{1 \times 1 \times H \times W}$. We compose the $Q_j > 0$ references in order:

$$\begin{aligned} C_j^{(0)} &= H_j, \\ C_j^{(q)} &= (1 - \alpha M_{jq}) \odot C_j^{(q-1)} + \alpha M_{jq} \odot P_{jq}, \quad q = 1, \dots, Q_j, \\ z_j^{\text{joint}} &= C_j^{(Q_j)}. \end{aligned} \tag{2}$$

Masks broadcast over channels and temporal frames, and $\alpha \in [0, 1]$ controls blending. We use $\alpha = 1$, so the last reference written wins in overlapping rectangles, and the canvas outside their union is unchanged. These are rectangular latent masks, not foreground masks: reference backgrounds may enter the placed regions. Appendix A.1 gives the resizing and margin rules.

The composed latent provides context for extracting reference features. It is neither emitted as a video frame nor pasted into the sampled output; the event chunk is generated from noise under its instruction. The boxes supply a spatial prior, rather than a guarantee of the final subject’s position. Contextualization changes the reference representation, so preserving existing content need not increase similarity to an isolated reference. New-ID and Existing-ID evaluate these objectives separately.

3.4 Cache Restoration after transient injection

We convert the joint latent into K/V with a clean DiT pass in a separate empty cache:

$$R_j = \Phi_\theta(z_j^{\text{joint}}, 0, \tilde{p}_j; \emptyset, \pi_{I_j}). \tag{3}$$

$\tilde{p}_j$ binds the available reference descriptions and layout to the current instruction. Φ_θ returns the K/V slices for the reference tokens in every layer. If each latent frame yields n_f tokens, the m frames supply $|I_j| = mn_f$ entries for the destination global slots I_j . The temporal coordinates π_{I_j} align the reference pass with those destination slots under the runtime’s RoPE convention. Spatial coordinates follow the composed canvas. The pass reads no running video K/V; its scene context comes from H_j .

Write $\mathcal{M}[I]$ for the K/V slices at physical cache indices I , across all layers, and $\eta(\mathcal{M})$ for metadata such as valid lengths, logical positions, and shot pointers. The operator $\text{Write}_I(\mathcal{M}, R)$ replaces only those slices with shape-matched R . Consider injection into an initialized stream, with cache $\mathcal{M}_j^-$ immediately before reference replacement, after any boundary preparation. The protected global slots I_j remain allocated through the update. For one injection chunk,

$$\begin{aligned} B_j &= \text{copy}(\mathcal{M}_j^-[I_j]), \\ \tilde{\mathcal{M}}_j &= \text{Write}_{I_j}(\mathcal{M}_j^-, R_j), \\ (\hat{z}_{k_j}, \mathcal{M}_j^{\text{den}}) &= \mathcal{D}_\theta(\epsilon_{k_j}, p_{k_j}; \tilde{\mathcal{M}}_j, c_{k_j}), \\ \overline{\mathcal{M}}_j &= \mathcal{U}_\theta(\mathcal{M}_j^{\text{den}}, \hat{z}_{k_j}, p_{k_j}; c_{k_j}), \\ \mathcal{M}_j^+ &= \text{Write}_{I_j}(\overline{\mathcal{M}}_j, B_j). \end{aligned} \tag{4}$$

The snapshot B_j owns its data. Current queries read the injected global entries together with shot and local history. Denoising is followed by a clean cache update, and only then are the saved entries restored. The next chunk starts from $\mathcal{M}_{k_j} = \mathcal{M}_j^+$. Conditional and unconditional caches follow the same transaction when classifier-free guidance is used. Nested writes are restored in reverse order.

By construction, restoration satisfies

$$\begin{aligned}\mathcal{M}_j^+[I_j] &= \mathcal{M}_j^-[I_j], \\ \mathcal{M}_j^+[\bar{I}_j] &= \bar{\mathcal{M}}_j[\bar{I}_j], \quad \eta(\mathcal{M}_j^+) = \eta(\bar{\mathcal{M}}_j),\end{aligned}\tag{5}$$

where $\bar{I}_j$ denotes all other allocated cache slots. Thus restoration preserves the clean update outside I_j and its advanced metadata; it does not roll back the stream. New subject information remains in generated history even after the direct reference K/V is removed. This is an invariant of the cache operation, not a guarantee against later identity drift or duplication.

Shot sink refresh is a separate policy within $\mathcal{U}_\theta$: Figure 2 uses S' for its resulting state, which can equal S . A refresh scheduled for a later chunk is applied there, not implicitly at injection. Hidden preroll and initialization are described in Appendix A.1. Removing Joint Latent replaces Equation 2 with full reference frames allocated to temporal slots. Removing Cache Restoration omits the final restore operation in Equation 4. These ablations distinguish reference construction from its duration in persistent memory, while retaining the measured tradeoff between new subject similarity and preservation.

3.5 ReferStream Bench for interactive subject injection

Conventional generation conditioned on references starts with its visual conditions already available. Interactive subject injection instead requires a generator to incorporate new appearance information after committing to a scene. Evaluating only whether a reference identity appears somewhere in the video misses a central requirement: it must appear at the requested stage while established subjects remain when the script requires them. ReferStream Bench therefore organizes each sequence around reference arrivals, evolving instructions, and expected subject presence.

Each entry pairs references with a chronological script specifying reveal boundaries, stage captions, normalized target regions where provided, and explicit cut flags. The frozen collection contains 48 InShot entries and 20 MultiShot entries. InShot requires introduction within a continuous shot; MultiShot permits a declared transition. The online protocol reveals conditions incrementally, while the offline comparison receives the references and script upfront. These information settings are reported separately. Requested departures are distinguished from unintended identity loss.

ReferStream Bench

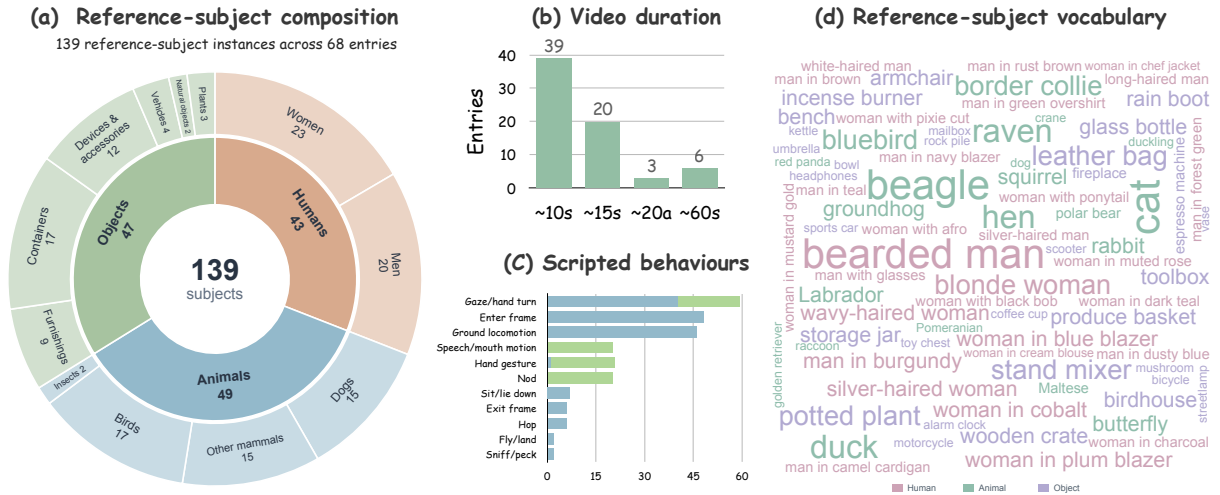


Figure 3 ReferStream Bench statistics. (a) Instances of reference subjects across the frozen collection; plants are included in objects. (b) Approximate durations of archived videos. (c) Scripted behavior frequencies, counted once per entry per label; labels can overlap. (d) Vocabulary of reference subjects, including recorded appearance descriptors. The figure describes benchmark content, not generation success or unique identities.

Figure 3 summarizes 139 instances of reference subjects: 43 humans, 49 animals, and 47 objects including plants. Sixty-five entries contain two reference subjects and three contain three. References are counted once per distinct path to a subject reference within an entry; reuse across entries is counted again. Approximate durations of 10, 15, 20, and 60 seconds occur in 39, 20, 3, and 6 entries, respectively, covering both short introductions and longer interactions after injection. The scripts include gaze/head turns, entry, locomotion, speech, gestures, and departure. Behavior labels overlap, and the vocabulary retains appearance descriptions such as clothing and hairstyle rather than only generic person labels.

Evaluation follows the event schedule: New-ID measures introduced subjects while present, Existing-ID measures established subjects that should persist, and Count checks the expected counts in each target category. Quality over the full video and matching against the active caption complement these subject measures. Allowed cuts and requested departures determine which observations are applicable. ReferStream Bench is a curated task resource rather than an exhaustive interaction taxonomy or an independently sampled held-out test set. Counting conventions and exclusions due to inconsistent reference assets are documented in Appendix A.2.

4 Experiments

4.1 Benchmark and comparison protocols

We evaluate on ReferStream Bench (Section 3.5) under InShot and MultiShot settings. All methods use the same frozen specifications and event schedules defined in physical time within their respective information settings. Entries with inconsistent reference assets are excluded uniformly across methods; the exclusions and temporal alignment rules are documented in Appendix A.

Online event conditioning. We compare ReferStream with LongLive 2.0 [5] under the same event schedule. The baseline receives stage captions and declared cuts but no reference images, including at initialization. ReferStream receives visual references at their scheduled reveals. The main comparison uses previously generated ReferStream outputs, aligned to the benchmark timeline as detailed in Appendix A.3. These comparisons evaluate the complete systems; paired ablations in Section 4.4 examine the contributions of individual components.

Offline conditioning on all references. VACE-14B [14], Phantom-14B [19], and HOMIE-14B [1] receive all references at time zero and the complete chronological script. Runs use seed 0, 50 sampling steps, and 832×480 output. VACE and Phantom use 16 fps and HOMIE uses 24 fps, with frame counts matched to the requested duration and text context extended where needed. LongLive and ReferStream videos use 1280×704 . Evaluation resizes all methods to 832×480 RGB and samples by physical time. These runs do not compare speed at matched resolutions.

Thin rules separate online methods (LongLive and ReferStream) from offline methods given all references (VACE, Phantom, and HOMIE). Offline methods have advance access to future references, so their scores describe adherence to the supplied script rather than adaptation to newly available information.

4.2 Main evaluation metrics

We report six complementary metrics without combining them into a weighted overall score. **Aesthetic quality** follows VBench [13], using CLIP ViT-L/14 [23] and the LAION linear aesthetic head [17]. **Motion smoothness** uses AMT-S [18] frame reconstruction, excluding triplets crossing declared cuts; a high score alone does not establish rich motion. **GME-T** uses GME’s multimodal embedding model [32], following the approach to matching frames and text in OpenS2V-Nexus [30], to compare each frame with its active stage caption. These columns summarize quality over the full video and text adherence. GME-T is not a direct test of each requested spatial relation or action.

New-ID measures reference similarity of the newly introduced subject after reveal, while it is expected to be present, pooling early and later observations into one score. **Existing-ID** measures subjects established before the event that should remain afterward. Grounding DINO detections [20], SAM 2.1 masks for each frame [24], and DINOv2-L features [22] provide one-to-one matching restricted to the same category; missing subjects score zero. **Count accuracy** requires exact agreement with the expected count in every target

Table 1 Overall comparison on ReferStream Bench using six computed metrics. Scores are raw values $\times 100$ ($\uparrow$), averaged with equal weight over eligible cases. Thin rules separate online event conditioning (above) from offline conditioning on all references (below). Bold marks the best point estimate in each column; the light blue row indicates our method. $\dagger$: automatic estimates that depend on detections.

Method	Aesth. $\uparrow$	Motion $\uparrow$	GME-T $\uparrow$	New-ID † $\uparrow$	Exist-ID † $\uparrow$	Count † $\uparrow$
LongLive-2.0	53.08	99.39	60.62	39.99	42.29	73.90
ReferStream (ours)	57.29	99.36	62.41	52.86	73.18	82.33
VACE-14B	57.57	99.23	57.83	43.33	65.97	51.22
Phantom-14B	48.37	97.10	50.90	29.23	62.72	31.07
HOMIE-14B	55.55	99.53	57.41	22.10	64.57	26.34

Table 2 Comparison by setting on ReferStream Bench. InShot evaluates subject introduction within a shot; MultiShot permits declared shot transitions. Scores are raw values $\times 100$ ($\uparrow$). Thin rules separate online event conditioning (above) from offline conditioning on all references (below). Bold marks the best point estimate in each setting; light blue rows indicate our method. $\dagger$: automatic estimates that depend on detections, subject to the limitations discussed in the text.

Setting	Method	Aesth. $\uparrow$	Motion $\uparrow$	GME-T $\uparrow$	New-ID † $\uparrow$	Exist-ID † $\uparrow$	Count † $\uparrow$
InShot	LongLive-2.0	51.93	99.40	65.18	49.21	48.62	72.70
	ReferStream (ours)	55.60	99.43	65.64	50.13	75.20	75.90
	VACE-14B	57.18	99.26	61.44	39.99	71.85	54.92
	Phantom-14B	46.23	96.96	53.36	23.64	63.65	23.06
	HOMIE-14B	53.96	99.51	60.88	12.60	69.87	19.67
MultiShot	LongLive-2.0	55.72	99.37	50.11	18.78	26.98	76.65
	ReferStream (ours)	61.17	99.21	54.99	59.13	68.28	97.13
	VACE-14B	58.46	99.16	49.53	51.00	51.74	42.69
	Phantom-14B	53.29	97.41	45.25	42.08	60.48	49.51
	HOMIE-14B	59.19	99.59	49.43	43.96	51.72	41.69

category. Categories outside the specified targets are ignored, and stages during which subjects depart are excluded from counting. Subject metrics average observations within events, events within cases, and cases with equal weight. All reported scores are multiplied by 100; ID columns remain similarity scores, not recognition percentages. Appendix A specifies checkpoints and preprocessing.

4.3 Overall results and comparison by setting

Overall comparison. Table 1 summarizes performance across ReferStream Bench, while Table 2 separates InShot and MultiShot. The overall scores use equal case weight rather than an unweighted average of the two setting means; each metric uses its eligible observations under the same rules for expected subject presence. ReferStream obtains the highest GME-T (62.41), New-ID (52.86), Existing-ID (73.18), and Count accuracy (82.33) among the compared methods. Relative to LongLive without visual references, the New-ID and Existing-ID gains are 12.87 and 30.89 score points, respectively, and Count accuracy increases by 8.43 percentage points. The stronger preservation gain is particularly relevant to interactive generation, where adding a new subject should retain established subjects that remain in the script.

The results for video quality reflect a different balance. ReferStream achieves 57.29 aesthetic quality, close to VACE’s 57.57, while HOMIE has the highest motion smoothness score (99.53). These dimensions are reported separately because a high smoothness score based on interpolation does not by itself establish identity fidelity or adherence to the changing instruction.

Subject introduction within a shot. InShot requires adapting an ongoing shot while maintaining existing scene content. In Table 2, ReferStream reaches 65.64 GME-T, 50.13 New-ID, 75.20 Existing-ID, and 75.90 Count accuracy. New-ID is close to LongLive without visual references (49.21), whereas Existing-ID is substantially higher than LongLive’s 48.62. Compared with offline methods given all references, ReferStream also has

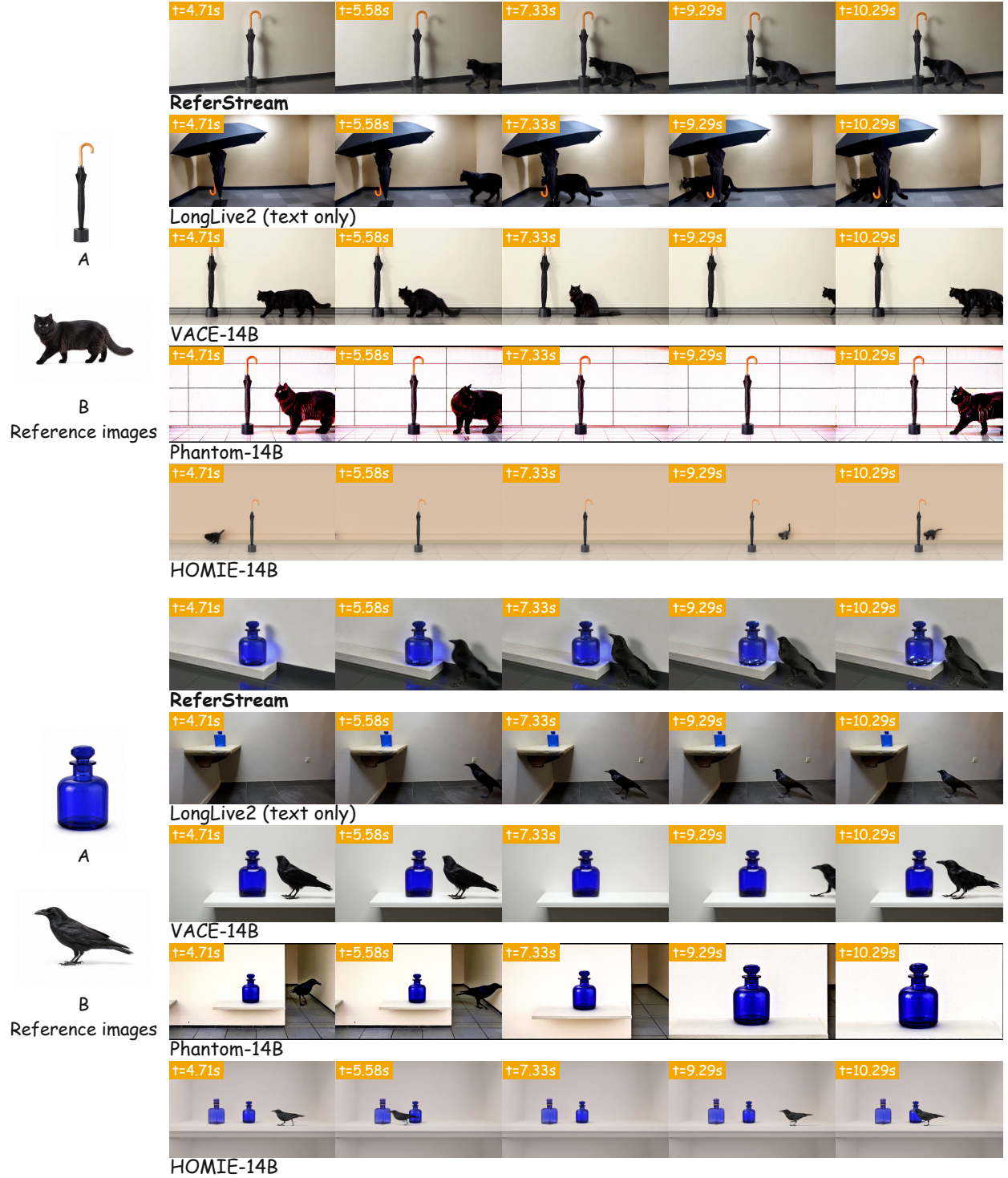


Figure 4 Qualitative comparisons on InShot: umbrella and black cat (top), cobalt bottle and raven (bottom). Reference A establishes the scene; B is revealed at $t = 5.21$ s. ReferStream introduces B into the ongoing shot while retaining A in the subsequent frames. Orange labels denote aligned benchmark time in seconds. LongLive2 receives online text events without visual references; VACE, Phantom, and HOMIE receive all references and the full script in advance. The selected frames illustrate reference fidelity, subject persistence, and count differences under these respective information settings.

higher identity and count scores. This pattern emphasizes preservation of the established scene’s subjects in addition to appearance of the new subject; it does not imply that every quality dimension improves, as VACE has higher aesthetic quality in this setting.

Subject introduction across shot transitions. MultiShot permits declared cuts and tests subject establishment and preservation under the corresponding stage instructions. ReferStream reaches 59.13 New-ID, 68.28 Existing-ID, and 97.13 Count accuracy, compared with 18.78, 26.98, and 76.65 for LongLive without visual references. It also obtains the highest GME-T (54.99) and aesthetic score (61.17) in this setting. The stronger separation from the baseline without visual references is consistent with the value of explicit visual references for subject appearance across shot transitions. These grouped results describe distinct benchmark settings rather than a controlled causal comparison between allowing and forbidding cuts.

Interpretation and evaluator limitations. The overall and grouped tables share the same scoring pipeline and set of valid reference assets. Offline methods receive all references and the complete script in advance, so their results reflect a different information setting from online event conditioning. ID and Count share automatic detections and can both be affected by missed subjects or category confusion. The fixed evaluator and aggregation audit support consistent measurement, but do not replace human validation or establish statistical significance.

4.4 Ablation studies

We compare full ReferStream (Full) with two variants that each remove one component. Without Joint Latent, the isolated reference is encoded as a full frame before the same K/V injection. Without Cache Restoration, the injected reference K/V remains in the global sink. Table 3 reports the common InShot comparison at seed 0. Table 4 evaluates Joint Latent under initial reference conditioning, with the cache policy held fixed. Scores average cases with equal weight on a 0–100 scale. Figures 5 and 6 complement these aggregate results with selected visual comparisons.

Table 3 Component ablations on the common InShot set (seed 0). Bold marks the best point estimate in each column; Full is shaded blue. Both components improve preservation and duplication point estimates at a cost in New-ID.

Variant	New-ID ↑	Existing-ID ↑	New Dup. ↓	Existing Dup. ↓
w/o Joint Latent	63.10	70.56	16.67	24.13
w/o Cache Restoration	52.94	69.36	20.49	29.17
ReferStream (Full)	51.31	72.17	14.41	23.26

Joint Latent. Constructing references within the current scene changes the balance between appearance matching and preservation. On the common set, Full raises Existing-ID from 70.56 to 72.17 and lowers new/existing duplication from 16.67/24.13 to 14.41/23.26, while New-ID decreases from 63.10 to 51.31. Figure 5 illustrates why these objectives should be considered together. Without joint construction, the red kettle changes shape after the woman appears; in the courtyard example, the bronze burner changes shape and an additional small vessel appears beside it. Full better retains the established objects in these selected sequences. The examples illustrate disruption of existing content that similarity to the newly introduced reference alone does not measure.

Table 4 Joint Latent under initial reference conditioning (seed 0). References are injected at generation start. Full and w/o Joint Latent use the same cache policy; only reference construction differs. Scores use the subset of short sequences, with metrics matching Table 3. Bold marks the best point estimate; Full is shaded blue.

Variant	New-ID ↑	Existing-ID ↑	New Dup. ↓	Existing Dup. ↓
w/o Joint Latent	64.23	74.55	13.68	21.79
ReferStream (Full)	52.22	75.30	12.82	20.09

Under initial reference conditioning (Table 4), Existing-ID rises from 74.55 to 75.30 and new/existing duplication falls from 13.68/21.79 to 12.82/20.09, while New-ID decreases from 64.23 to 52.22. Here references are injected once at generation start, so this comparison tests reference construction without an already emit-

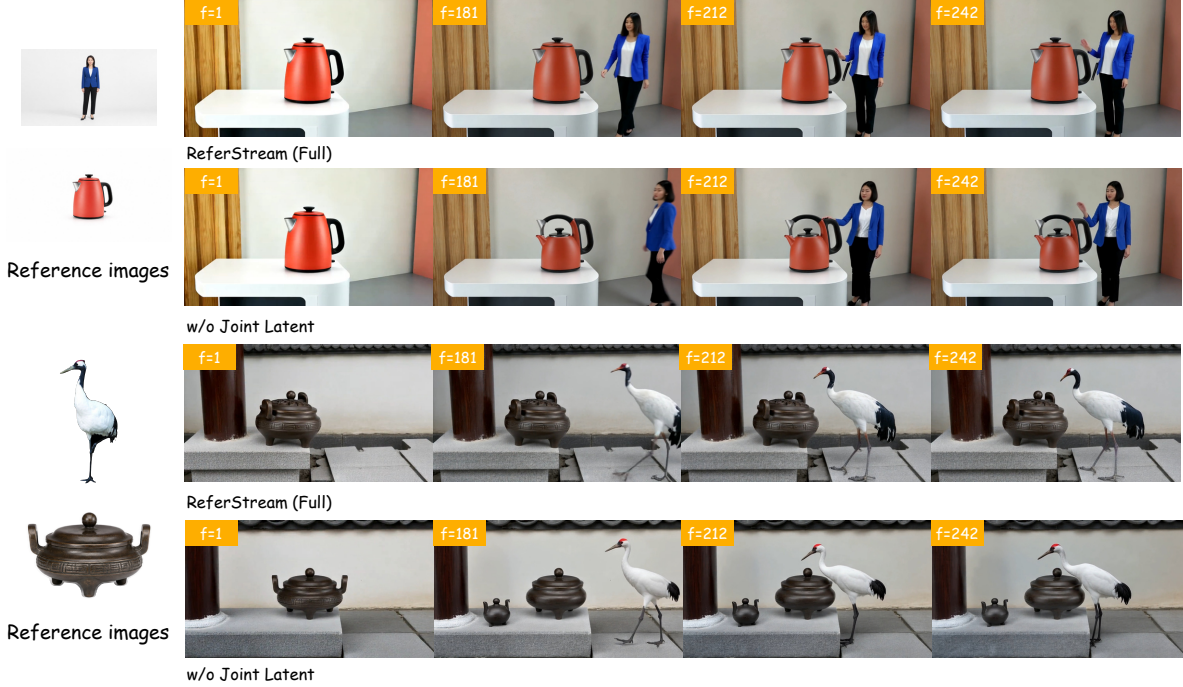


Figure 5 Effect of Joint Latent on established subjects. Full and w/o Joint Latent are compared on the kettle and woman example (top) and the burner and crane example (bottom). Without joint construction, the kettle and burner change shape, and an extra vessel appears beside the burner. Reference images are shown on the left; f denotes the source frame index.

ted prefix. These small preservation and duplication changes are treated as single-seed trends; the New-ID decrease remains a measured cost.

Cache Restoration. Cache Restoration recovers the original global entries after the injection chunk has been committed, leaving the generated history available to subsequent chunks. On the common set, it raises Existing-ID from 69.36 to 72.17 and lowers new/existing duplication from 20.49/29.17 to 14.41/23.26, at a smaller New-ID decrease from 52.94 to 51.31.

Figure 6 shows two ways the variants diverge after the new subject appears. Without restoration, bright star patterns progressively cover the bag-dog scene; in the mixer-duck scene, an additional duck appears on the left. Full avoids these particular failures in the displayed frames. These observations illustrate later scene disruption and unwanted instances, rather than establishing their frequency across the benchmark.

The separate multistage comparison in Appendix B more directly examines preservation over successive events. Cache Restoration raises Existing-ID from 46.44 to 58.60. New-ID is lower and new duplication is higher on that subset, so restoration does not improve every subject metric. Together, the ablations support distinct roles for the two components: Joint Latent supplies scene context when constructing reference features, whereas Cache Restoration limits how long those features occupy persistent memory. The tables quantify the resulting tradeoffs; the selected images illustrate their visual consequences.

4.5 Qualitative comparisons

Figure 4 compares two InShot sequences at aligned benchmark times. In each example, generation first establishes subject A; a new reference B is then supplied during the ongoing shot. The interface permits the user to choose when to submit references, with each arrival applied at the next chunk boundary (Section 3.2). The new subject is integrated through continued generation while the emitted prefix remains unchanged. Here, B is revealed at $t = 5.21$ s, between the first and second displayed columns.

In the upper example, ReferStream retains the closed umbrella and its orange handle as the black cat enters

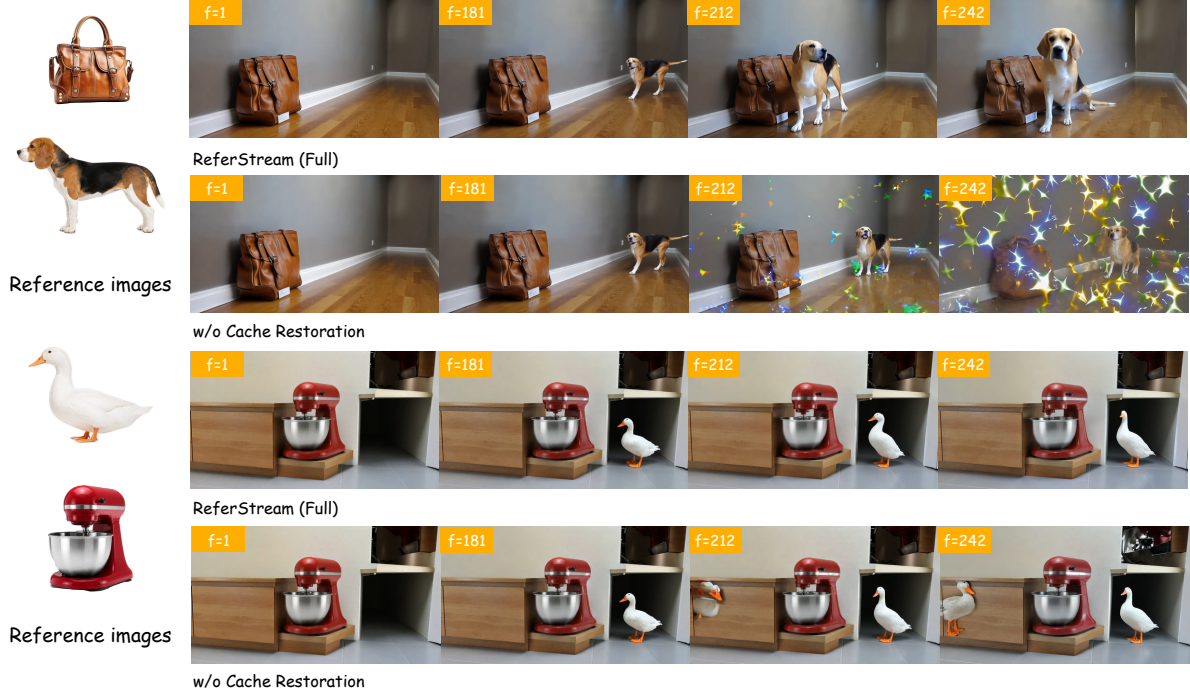


Figure 6 Effect of Cache Restoration after subject injection. Full and w/o Cache Restoration are compared on the bag and dog example (top) and the mixer and duck example (bottom). Without restoration, star patterns progressively cover the first scene and an extra duck appears in the second. Reference images are shown on the left; f denotes the source frame index.

from the right and remains beside the stand. LongLive2 follows the cat instruction, but its open umbrella differs from reference A even before the event. In the lower example, ReferStream preserves the cobalt bottle’s square body and rounded stopper while the raven joins it on the shelf. HOMIE produces two bottles, and Phantom loses the raven in the later displayed frames. These examples illustrate the joint requirement of introducing the new subject and retaining the appearance and number of established subjects. They complement the identity and count results in Tables 1 and 2; Appendix C provides another InShot example and a MultiShot comparison.

5 Conclusion

We introduced ReferStream for injecting newly supplied visual references into a video during causal streaming generation. Joint Latent constructs reference features in the current scene, while Cache Restoration limits their direct influence to the injection chunk and retains the generated history. Together, they allow a frozen generator to incorporate new subjects without revising emitted frames. ReferStream Bench evaluates this interaction through reference arrival schedules and explicit expectations for new and established subjects. Across InShot and MultiShot, ReferStream achieves higher measured text adherence, subject similarity, and count accuracy than the compared baselines, with mixed results for video quality. The ablations show that preserving established content can come at a cost in similarity to the new reference. These findings support considering both the scene context used to construct reference features and the duration of their influence in persistent memory. Broader evaluation across seeds and further validation of automatic subject metrics remain useful next steps.

References

- [1] Cai, Y., Chen, N., Xie, R., Pan, J., Jiang, C., Chen, C., Zhou, W., Sun, Z., Xue, W., Luo, W., and Guo, Y. Homie: Human-object centric video personalization via multimodal intelligent enhancement, 2026. URL <https://arxiv.org/abs/2607.18217>.
- [2] Chen, B., Monsó, D. M., Du, Y., Simchowitz, M., Tedrake, R., and Sitzmann, V. Diffusion forcing: Next-token prediction meets full-sequence diffusion. In *Advances in Neural Information Processing Systems*, volume 37, pp. 24081–24125, 2024. URL <https://arxiv.org/abs/2407.01392>.
- [3] Chen, G., Lin, D., Yang, J., Lin, C., Zhu, J., Fan, M., Zhang, H., Chen, S., Chen, Z., Ma, C., et al. SkyReels-V2: Infinite-length Film Generative Model. *arXiv preprint arXiv:2504.13074*, 2025. URL <https://arxiv.org/abs/2504.13074>.
- [4] Chen, T.-S., Siarohin, A., Menapace, W., Fang, Y., Lee, K. S., Skorokhodov, I., Aberman, K., Zhu, J.-Y., Yang, M.-H., and Tulyakov, S. Multi-subject Open-set Personalization in Video Generation. *arXiv preprint arXiv:2501.06187*, 2025. URL <https://arxiv.org/abs/2501.06187>.
- [5] Chen, Y., Wang, L., Huang, W., Yang, S., Zhang, B., Xiao, Y., Chu, R., Mao, W., Hu, Q., Liu, S., Zhao, Y., Mao, H., Chen, Y.-C., Xie, E., Qi, X., and Han, S. Longlive2.0: An nvfp4 parallel infrastructure for long video generation. *arXiv preprint arXiv: 2605.18739*, 2026.
- [6] Deng, Y., Yin, Y., Guo, X., Wang, Y., Fang, J. Z., Yuan, S., Yang, Y., Wang, A., Liu, B., Huang, H., et al. MAGREF: Masked Guidance for Any-Reference Video Generation with Subject Disentanglement. *arXiv preprint arXiv:2505.23742*, 2025. URL <https://arxiv.org/abs/2505.23742>.
- [7] Feng, T., Li, Z., Yang, S., Xi, H., Li, M., Li, X., Zhang, L., Yang, K., Peng, K., Han, S., et al. StreamDiffusionV2: A Streaming System for Dynamic and Interactive Video Generation. In *Proceedings of Machine Learning and Systems*, 2026. URL <https://arxiv.org/abs/2511.07399>.
- [8] Girish, S., Ivanov, V., Chen, T.-S., Chen, H., Siarohin, A., and Tulyakov, S. AlcheMinT: Fine-grained Temporal Control for Multi-Reference Consistent Video Generation. *arXiv preprint arXiv:2512.10943*, 2025. URL <https://arxiv.org/abs/2512.10943>.
- [9] HaCohen, Y., Chiprut, N., Brazowski, B., Shalem, D., Moshe, D., Richardson, E., Levin, E., Shiran, G., Zabari, N., Gordon, O., et al. LTX-Video: Realtime Video Latent Diffusion. *arXiv preprint arXiv:2501.00103*, 2024. URL <https://arxiv.org/abs/2501.00103>.
- [10] HaCohen, Y., Brazowski, B., Chiprut, N., Bitterman, Y., Kvochko, A., Berkowitz, A., Shalem, D., Lifschitz, D., Moshe, D., Porat, E., et al. LTX-2: Efficient Joint Audio-Visual Foundation Model. *arXiv preprint arXiv:2601.03233*, 2026. URL <https://arxiv.org/abs/2601.03233>.
- [11] Huang, X., Li, Z., He, G., Zhou, M., and Shechtman, E. Self Forcing: Bridging the Train-Test Gap in Autoregressive Video Diffusion. In *Advances in Neural Information Processing Systems*, 2025. URL <https://arxiv.org/abs/2506.08009>.
- [12] Huang, Y., Yuan, Z., Liu, Q., Wang, Q., Wang, X., Zhang, R., Wan, P., Zhang, D., and Gai, K. ConceptMaster: Multi-Concept Video Customization on Diffusion Transformer Models Without Test-Time Tuning. *arXiv preprint arXiv:2501.04698*, 2025. URL <https://arxiv.org/abs/2501.04698>.
- [13] Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., Wang, Y., Chen, X., Wang, L., Lin, D., Qiao, Y., and Liu, Z. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.
- [14] Jiang, Z., Han, Z., Mao, C., Zhang, J., Pan, Y., and Liu, Y. VACE: All-in-one video creation and editing. *arXiv preprint arXiv:2503.07598*, 2025. URL <https://arxiv.org/abs/2503.07598>.
- [15] Kodaira, A., Xu, C., Hazama, T., Yoshimoto, T., Ohno, K., Mitsuhori, S., Sugano, S., Cho, H., Liu, Z., Tomizuka, M., et al. StreamDiffusion: A Pipeline-level Solution for Real-time Interactive Generation. *arXiv preprint arXiv:2312.12491*, 2023. URL <https://arxiv.org/abs/2312.12491>.
- [16] Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al. HunyuanVideo: A Systematic Framework For Large Video Generative Models. *arXiv preprint arXiv:2412.03603*, 2024. URL <https://arxiv.org/abs/2412.03603>.
- [17] LAION-AI. LAION-Aesthetics Predictor V1. Software and pretrained linear aesthetic weights, n.d. URL <https://github.com/LAION-AI/aesthetic-predictor>. ViT-L/14 checkpoint: sa_0_4_vit_l_14_linear.pth. Accessed September 22, 2026.

- [18] Li, Z., Zhu, Z.-L., Han, L.-H., Hou, Q., Guo, C.-L., and Cheng, M.-M. AMT: All-pairs multi-field transforms for efficient frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9801–9810, 2023. URL <https://arxiv.org/abs/2304.09790>.
- [19] Liu, L., Ma, T., Li, B., Chen, Z., Liu, J., Li, G., Zhou, S., He, Q., and Wu, X. Phantom: Subject-consistent video generation via cross-modal alignment. *arXiv preprint arXiv:2502.11079*, 2025.
- [20] Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., Zhu, J., and Zhang, L. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. URL <https://arxiv.org/abs/2303.05499>.
- [21] MiniMax. MiniMax H3. Official model release, 2026. URL <https://www.minimax.cn/blog/minimax-h3>.
- [22] Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jégou, H., Mairal, J., Labatut, P., Joulin, A., and Bojanowski, P. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. URL <https://arxiv.org/abs/2304.07193>.
- [23] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021. URL <https://proceedings.mlr.press/v139/radford21a.html>.
- [24] Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K. V., Carion, N., Wu, C.-Y., Girshick, R., Dollár, P., and Feichtenhofer, C. SAM 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. URL <https://arxiv.org/abs/2408.00714>.
- [25] Wan Team, Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.-W., Chen, D., Yu, F., Zhao, H., Yang, J., et al. Wan: Open and Advanced Large-Scale Video Generative Models. *arXiv preprint arXiv:2503.20314*, 2025. URL <https://arxiv.org/abs/2503.20314>.
- [26] Yang, S., Huang, W., Chu, R., Xiao, Y., Zhao, Y., Wang, X., Li, M., Xie, E., Chen, Y., Lu, Y., et al. Longlive: Real-time interactive long video generation. In *ICLR*, 2026.
- [27] Yang, Y., Zhang, T., Huang, W., Chen, J., Wu, B., He, X., Cai, D., Li, B., and Jiang, P.-T. Anchor Forcing: Anchor Memory and Tri-Region RoPE for Interactive Streaming Video Diffusion. *arXiv preprint arXiv:2603.13405*, 2026. URL <https://arxiv.org/abs/2603.13405>.
- [28] Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al. CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer. *arXiv preprint arXiv:2408.06072*, 2024. URL <https://arxiv.org/abs/2408.06072>.
- [29] Yin, T., Zhang, Q., Zhang, R., Freeman, W. T., Durand, F., Shechtman, E., and Huang, X. From Slow Bidirectional to Fast Autoregressive Video Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL <https://arxiv.org/abs/2412.07772>.
- [30] Yuan, S., He, X., Deng, Y., Ye, Y., Huang, J., Lin, B., Luo, J., and Yuan, L. Opens2v-nexus: A detailed benchmark and million-scale dataset for subject-to-video generation. *arXiv preprint arXiv:2505.20292*, 2025.
- [31] Zhang, L., Cai, S., Li, M., Wetzstein, G., and Agrawala, M. Frame Context Packing and Drift Prevention in Next-Frame-Prediction Video Diffusion Models. *arXiv preprint arXiv:2504.12626*, 2025. URL <https://arxiv.org/abs/2504.12626>.
- [32] Zhang, X., Zhang, Y., Xie, W., Li, M., Dai, Z., Long, D., Xie, P., Zhang, M., Li, W., and Zhang, M. GME: Improving universal multimodal retrieval by multimodal llms. *arXiv preprint arXiv:2412.16855*, 2024. URL <https://arxiv.org/abs/2412.16855>.

A Implementation and evaluation details

A.1 Reference construction and cache updates

The reference branch uses the frozen Wan VAE and LongLive2 DiT. The exported configuration generates eight latent frames per chunk, uses eight global sink frames and four sampling steps, and configures the local attention window for 32 latent frames. Reference K/V is stored without quantization. The backbone’s quantization and parallelization options are separate from this cache intervention.

For a normalized target box, Joint Latent insets each side by $\delta = 0.05$, clips the coordinates to $[0, 1]$, and rejects collapsed boxes. The endpoints are scaled to latent resolution, rounded, and clamped to a nonempty grid rectangle. The full reference encoding is bilinearly resized to this rectangle with `align_corners=False`; explicit boxes override layout defaults. The latest clean scene latent is repeated across reference slots, and the full configuration uses unit reference strength. With no generated history, the canvas is zero. Multiple rectangles are written in reference order. Without Joint Latent, full reference latents occupy temporal slots, with separate blocks allocated to multiple references.

A clean reference pass runs in an empty temporary cache with temporal coordinates aligned to the destination global slots, starting at slot zero. Original K/V tensors are cloned before replacement in every transformer layer. Denoising writes temporary current-chunk K/V; the clean pass replaces these writes before Cache Restoration restores the saved global entries. The generated latent, other cache entries, and advanced indices remain. Overlapping temporary writes are restored in reverse order. When classifier-free guidance is enabled, the transaction applies to both conditional and unconditional caches.

The runtime can also pin a generated chunk as a refreshed shot anchor at a configured delay after the reference event. This policy is determined by the archived configuration and remains part of $\mathcal{U}_\theta$; Equation 5 restores the global entries without requiring the entire shot cache to remain unchanged. Hidden preroll establishes the initial scene in the archived InShot schedules. Later references have their own event intervals. Text updates refresh text conditioning without automatically cutting the shot or discarding its visual history.

A.2 Benchmark curation

The frozen collection contains 68 entries and 67 base scenario names: the storeroom cat/jar scenario has two archived variants. Reference statistics deduplicate subject paths within each entry, exclude environment-only references, and count reuse across entries. Plants belong to objects; vocabulary labels retain appearance descriptions and do not denote unique identities. Approximate 10/15/20/60 second duration bins contain 39/20/3/6 entries, based on archived video metadata. Behavior labels are counted once per entry, can overlap, and include hopping under ground locomotion. The collection was frozen before baseline evaluation and is a curated development resource.

A.3 Output sources and temporal alignment

The evaluator maps decoded source frames to the canonical timeline without pixel interpolation or video recompression. For 24 fps presentation exports with one leading frame removed, the first retained frame supplies the missing canonical position. Legacy 16 fps MultiShot outputs use recorded leading-frame corrections and nominal five-second stages; LongLive maps four chunks to each such stage. For edited 30 fps long videos, endpoint-preserving indices $j = \text{round}(i(N_s - 1)/(N_t - 1))$ map target frames to source frames, where N_s and N_t are the source and target frame counts. Where edit positions are unrecorded, this mapping assumes uniform progression through the source clip and does not recover exact local event boundaries.

A.4 Evaluation pipeline and pretrained components

RGB frames are resized to 832×480 with Lanczos before evaluator preprocessing. All methods use the same evaluators and thresholds. Table 5 lists the components; source URLs, repository revisions where available, and checkpoint hashes are retained in the experiment records.

Video quality and text adherence. Aesthetic and GME-T samples use $t = 0.125 + 0.25n$ seconds, taking the nearest native frame with ties resolved forward. Aesthetic quality uses normalized CLIP ViT-L/14 features and the LAION linear predictor, divided by ten. GME-T computes similarity between each frame and its active caption using the 7B GME checkpoint in half precision. Motion evaluation uses 8 fps samples, triplets

with a stride of two, and one minus normalized AMT-S reconstruction error; triplets crossing declared cuts are excluded. Each metric is averaged within a video and then across cases with equal weight.

Subject similarity and counts. Grounding DINO Tiny queries the target categories jointly, using a rodent alias for groundhog. Detections are scored at threshold 0.50, with within-category NMS at 0.5 IoU and cross-category NMS at 0.7. Candidates from 0.20 are retained for inspection. SAM 2.1 Large selects the highest-confidence mask for each box. Pixels outside the mask are replaced with gray (RGB 127), and the box crop is encoded with DINOv2-L without registers. References receive the same preprocessing. Hungarian matching is restricted to the same category; missing subjects score zero and cosine similarities are clipped to $[0, 1]$. ID uses continuous similarities, without an identity threshold. Scheduled departure transitions are excluded from ID and Count. Subject scores average observations within events, events within cases, and then cases equally.

Table 5 Evaluation components and pretrained checkpoints. Linked names point to the checkpoint provider or official implementation.

Role	Checkpoint / implementation	Source
Aesthetic features	CLIP ViT-L/14	Radford et al. [23]
Aesthetic head	LAION V1 ViT-L/14 linear head	LAION-AI [17]
Motion interpolation	AMT-S	Li et al. [18]
Text adherence	gme-Qwen2-VL-7B-Instruct	Zhang et al. [32]
Object detection	Grounding DINO Tiny	Liu et al. [20]
Foreground masking	SAM 2.1 Hiera Large	Ravi et al. [24]
Subject features	DINOv2 Large (ViT-L/14)	Oquab et al. [22]

Scope of the automatic evaluation. The detector was adjusted on a fixed development sample reviewed with model assistance before inspecting aggregate rankings, then kept unchanged across methods. No method-specific thresholds or manual score corrections are used. Audits checked aggregation against stored observations. Chicken/groundhog confusion, missed jars, and partial detections remain possible, so these checks do not replace independent human validation. The event windows, presence rules, and aggregation are task-specific adaptations of the cited evaluators. The reported comparisons do not establish robustness across generation seeds.

B Additional ablations

B.1 Subset design and interpretation

The common comparison uses 48 paired InShot cases at seed 0. The initialization subset contains 39 ten-second cases with a single reference injection at generation start. The multistage subset contains nine cases: three lasting 15 seconds and six lasting 60 seconds. Membership follows the archived schedules, independently of metric outcomes. Multistage denotes successive events within these schedules, rather than the MultiShot setting.

Full denotes ReferStream with both components. The initialization comparison holds the cache policy fixed and changes only Joint Latent (Table 4); the multistage comparison changes only Cache Restoration (Table 6). Full means differ because the subsets contain different cases. Their absolute values should not be compared with one another or with the separate main experiment archive.

New-ID and Existing-ID measure reference similarity, which may reflect category, appearance, and context as well as identity. At initialization, Existing-ID labels the other expected subject; it does not test preservation across a later reference arrival. Duplication measures unwanted copies of each subject and is distinct from the main Count metric, which requires all target counts to match. The single-seed comparisons describe the observed tradeoffs rather than establish statistical significance.

B.2 Cache Restoration over successive events

Table 6 tests whether restoring the original global memory helps retain subjects through later events. Existing-ID rises from 46.44 to 58.60 and Existing Duplication falls from 46.30 to 37.04. New-ID decreases from 53.67

to 47.38, while New Duplication increases from 16.67 to 21.30. This pattern supports a preservation benefit on the multistage subset, together with a cost to the newly introduced subject. It complements the common-set ablation and concerns a different temporal setting from initial reference conditioning.

Table 6 Cache Restoration on the multistage subset (seed 0). Full restores the overwritten global entries; w/o Cache Restoration retains raw reference K/V. Bold marks the best point estimate; Full is shaded blue.

Variant	New-ID ↑	Existing-ID ↑	New Dup. ↓	Existing Dup. ↓
w/o Cache Restoration	53.67	46.44	16.67	46.30
ReferStream (Full)	47.38	58.60	21.30	37.04

C Additional qualitative comparisons

The following cases complement Figure 4. Each row contains five native frames at common benchmark times, using the alignment in Appendix A. References appear on the left. LongLive2 receives online text events; VACE, Phantom, and HOMIE receive all references and the script in advance.

C.1 Preserving an object during InShot injection

In Figure 7, the white hen reference arrives after a bronze burner has been established. The burner should remain as the hen enters and stays nearby. This case tests whether a new reference can be integrated while preserving an existing object’s structure.



Figure 7 InShot injection of a white hen beside a bronze burner. Reference B arrives at $t = 5.21$ s; A should remain in the same shot. Orange labels indicate aligned benchmark time in seconds.

ReferStream retains the burner’s low rounded body, lid, and handles as the hen enters on its right. In the selected baseline frames, LongLive2 shows a different burner shape from the outset; Phantom and HOMIE do not retain the hen in the later views. VACE shows the hen in the final view but omits it in the middle columns. These frames illustrate why the introduced subject’s appearance and preservation of existing content must be considered together.

C.2 Controlling composition across MultiShot stages

Figure 8 follows a furniture design discussion with the requested order $A \rightarrow B \rightarrow A + B$. The five columns sample the A stage, two moments in the B stage, and two moments in the joint discussion. This case tests subject composition across declared shot transitions.



Figure 8 MultiShot subject composition in a furniture design discussion. The requested composition is $A, B, B, A + B, A + B$ across the displayed columns. Orange labels indicate aligned benchmark time in seconds.

ReferStream follows the stage changes while retaining charcoal clothing with a copper accent for A and blue clothing for B. LongLive2 still shows A during the requested B stage. VACE and HOMIE frequently show both people in individual stages, and Phantom does so in several views, with a pale transition at $t = 5.88$ s. These are failures to match the requested composition under the stated information settings; early joint appearances by offline methods do not imply access to unavailable references.

Together, the InShot and MultiShot examples illustrate two uses of the same interface: introducing a subject within an ongoing shot and updating composition across a declared cut. Users may submit references at arbitrary times, with each arrival applied at the next chunk boundary. The benchmark fixes these times for reproducible comparisons, while the generator conditions on each reference only after its arrival.