# CHRONOGUIDE: TRAINING-FREE TEMPORAL CONTROL FOR VIDEO GENERATION VIA CROSS-STATE ATTENTION

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

## ABSTRACT

Despite major gains in visual quality, text-to-video models remain unreliable at precise temporal control. Objects may appear early or late, persist too long, or fail to return on time, and actions may peak off cue. We identify a structural source of these failures in spatiotemporal self-attention. The same cross-frame communication that supports coherence also transfers features between frames requiring different states, such as presence and absence, pulling them away from the requested schedule. ChronoGuide addresses this conflict with a training-free *cross-state attention mechanism*. Its core operator preserves live attention weights and same-state contributions, replacing only values crossing state boundaries with matched features computed under the receiving frame's requested state. This targets cross-state interference while retaining same-state communication, without training or iterative latent optimization. We further introduce a mixed-lifecycle object evaluation covering entry, exit, bounded presence, and return, and a complementary evaluation using an open-vocabulary detector. Across all benchmark panels evaluated against TempoControl, ChronoGuide improves every reported metric. On the standard 80-prompt one-object Wan benchmark, it reaches 90.06% temporal accuracy, 83.88% presence accuracy, and 96.25% absence accuracy versus TempoControl's 83.56%, 79.75%, and 87.38%. It also uses substantially less peak VRAM than TempoControl, enabling deployment on lower-memory GPUs and batched inference. ChronoGuide transfers across continuous and discrete video diffusion models, including CogVideoX and URSA.

## 1 INTRODUCTION

Video generation has progressed from early adversarial models of scene dynamics and motion (Vondrick et al., 2016) to modern text-to-video systems based largely on latent diffusion and flow matching (Ho et al., 2022b; Wan et al., 2025). Autoregressive and masked-token models offer an alternative formulation (Kondratyuk et al., 2024; Villegas et al., 2023), with discrete diffusion recently extending this direction to video (Deng et al., 2026). As visual quality has improved, control has expanded to layout, trajectories, object motion, and camera motion (Wang et al., 2023; 2024d).

Precise temporal control of a *controlled concept*, an object or action, remains substantially harder. Objects require presence/absence timing, while actions require peak-motion timing. Existing methods mainly improve temporal conditioning or attention alignment. MinT learns time-aware text–video interactions through fine-tuning (Wu et al., 2025), whereas training-free methods such as SwitchCraft and TS-Attn steer cross-attention at inference (Xu et al., 2026; Zhang et al., 2026). TempoControl (TC) keeps model weights frozen but repeatedly optimizes the latent variables during denoising (Schiber et al., 2026). These approaches primarily shape temporal conditioning or attention allocation, but do not directly control what video frames communicate to one another.

This distinction matters because spatiotemporal self-attention can bleed temporal-state information across frames. Frames assigned different states exchange features, potentially shifting one toward the other's state; for example, present-frame features can drive an absent-frame toward presence. Yet severing these connections would disrupt the cross-frame communication needed for coherent video. Hence, we view temporal control as both a conditioning and a communication problem.



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




Figure 1: ChronoGuide delays appearance toward the benchmark's 4.25 s onset.

We introduce **ChronoGuide (CG)**, a training-free, forward-only method that addresses this communication directly. A framewise schedule defines each controlled concept's requested state. CG computes state-conditioned value reads from the same current video and, when frames assigned different states interact, substitutes the value matching the state requested by the frame being updated. Thus, CG changes the transmitted content while preserving live attention allocation, requiring neither fine-tuning nor expensive iterative latent optimization (Figure 1).

We evaluate temporal adherence separately from visual quality. VBench mainly measures the latter (Huang et al., 2024). For temporal adherence, we retain TC's YOLOv10 object-timing and optical-flow movement protocols (Schiber et al., 2026; Wang et al., 2024a), and add OWLv2 to assess robustness to detector choice (Minderer et al., 2023). Optical flow measures when motion peaks rather than action identity; tested semantic evaluators failed our reliability checks (Appendix I). We also evaluate mixed lifecycles covering entry, exit, bounded presence, and return.

Our contributions are threefold:

- **Cross-state attention for training-free temporal control.** We identify cross-state feature propagation as a source of temporal control errors and introduce, to our knowledge, the first mechanism that directly corrects values transmitted across temporal-state boundaries while preserving attention weights and same-state communication.
- **Stronger temporal control at lower inference cost.** CG improves every headline benchmark metric over same-host TC, raising temporal accuracy from 83.56% to 90.06% for one-object control and from 53.17% to 67.83% for two-object control. It uses substantially less peak VRAM than TC, and transfers across continuous and discrete video models.
- **Mixed lifecycles and open-vocabulary evaluation.** We introduce temporal evaluation covering entry, exit, bounded presence, and return, and extend closed-vocabulary protocols with OWLv2. Our detector analysis reveals differences between YOLOv10 and OWLv2 presence judgments, motivating complementary open-vocabulary evaluation.

## 2 Related Work

**Video generators and temporal communication.** Modern video generators combine latent diffusion, Transformer denoisers, and flow matching to model spatiotemporal interactions (Rombach et al., 2022; Ho et al., 2022b; Singer et al., 2023; Ho et al., 2022a; Blattmann et al., 2023b;a; Peebles & Xie, 2023; Lipman et al., 2023; Liu et al., 2023; Guo et al., 2024; Chen et al., 2024; Bar-Tal et al., 2024; Gupta et al., 2024; Skorokhodov et al., 2024; Polyak et al., 2025; Jin et al., 2025; Kong et al., 2025; Wan et al., 2025; Yang et al., 2025b; Zheng et al., 2026; HaCohen et al., 2026).

**Temporal control.** Video control has largely focused on layouts, trajectories, object motion, cameras, and 3D constraints (Wang et al., 2023; Zhang et al., 2024a; Wang et al., 2024d; Yang et al., 2024; Wang et al., 2024b; Ma et al., 2024; Shi et al., 2024; Geng et al., 2025; Lei et al., 2025; Zhang et al., 2025; Ji et al., 2025; Li et al., 2025; He et al., 2025; Bahmani et al., 2025; Ren et al., 2025; Ling et al., 2025; Singer et al., 2026). Temporal structure is introduced through masks, layouts, or structured prompts (Jain et al., 2024; Tian et al., 2024; Lian et al., 2024; Lin et al., 2024; Lo et al., 2025; Feng et al., 2025), multi-prompt and multi-scene generation (Qiu et al., 2024; Oh et al., 2024; Qi et al., 2025), learned or encoded temporal intervals (Wu et al., 2025; Girish et al., 2026; Shu

et al., 2026), and training-free attention or prompt steering (Cai et al., 2025; Xu et al., 2026; Schiber et al., 2026; Chen et al., 2026; Liao et al., 2026; Zhang et al., 2026). MinT and AlcheMinT encode temporal assignments through learned conditioning; TIE represents event intervals inside cross-attention; TC optimizes latents from attention-derived objectives; SwitchCraft steers frame queries; and TS-Attn, TunerDiT, and Prompt Relay rearrange, mask, or route event conditioning. In contrast, CG targets the feature content transmitted between frames with different requested temporal states, thereby limiting cross-state interference.

**Feature transfer and inference-time guidance.** Inference-time control includes compositional, scheduled, and self-guidance (Ho & Salimans, 2021; Dhariwal & Nichol, 2021; Liu et al., 2022; Brack et al., 2023; Sadat et al., 2024; Kynkäänniemi et al., 2024; Karras et al., 2024; Ahn et al., 2024; Hyung et al., 2025; Ezra et al., 2025; Zhou et al., 2026), loss-guided and frame-level sampling (Chung et al., 2023; Yu et al., 2023; Song et al., 2023; Bansal et al., 2023; He et al., 2024b; Ye et al., 2024; Jang et al., 2026), guidance for video and discrete generators (Lee et al., 2025; Kim et al., 2025; Shaulov et al., 2025; Nisonoff et al., 2025; Rojas et al., 2026), and attention or feature manipulation (Hertz et al., 2023; Chefer et al., 2023; Zhang et al., 2024b; Dahary et al., 2024; Tumanyan et al., 2023; Cao et al., 2023; Bar-Tal et al., 2023; Mo et al., 2024; Mou et al., 2024; Tewel et al., 2024; Zhou et al., 2024). Text2Video-Zero, TokenFlow, and FateZero modify or propagate information across frames for consistency and editing (Khachatryan et al., 2023; Geyer et al., 2024; Qi et al., 2023). DiTCtrl uses mask-guided attention sharing across prompt-conditioned segments (Cai et al., 2025), while TPD constructs an early-segment counterfactual in CFG space to restore suppressed late-event signals (Kang & Zwicker, 2026). CG computes alternative values from the same current video under the receiver's requested state. These replace only cross-state values, preserving live attention weights and same-state contributions.

**Temporal evaluation.** General benchmarks measure video quality, consistency, composition, and dynamics (Unterthiner et al., 2019; Huang et al., 2024; Liu et al., 2024b; Liao et al., 2024; He et al., 2024a; Sun et al., 2025; Han et al., 2025; Liu et al., 2025; Yang et al., 2025a). Temporal visibility and motion are commonly evaluated with object detectors, vision–language models, and optical flow (Wang et al., 2024a; Minderer et al., 2022; 2023; Radford et al., 2021; Liu et al., 2024a; Teed & Deng, 2020; Xu et al., 2022; Huang et al., 2022). We retain TC's YOLO-based object timing and optical-flow movement protocols for direct comparison, add OWLv2 for more detector robustness, and introduce mixed lifecycles covering entry, exit, bounded presence, and return.

# 3 Background

## 3.1 Conditional Video Generation as Iterative Refinement

Let $x \in \mathbb{R}^{T\times H\times W\times C}$ denote a rendered video with $T$ frames, spatial resolution $H \times W$, and $C$ channels, and let $p$ denote its text prompt. A pretrained generator refines an internal state $s_\tau$ over steps $\tau = N, \dots, 1$, with model time $t_\tau \in [0, 1]$ set by the inference schedule.

At step $\tau$, a frozen network $f_\theta$ predicts the model-specific quantity used to update the current state. When classifier-free guidance (CFG) (Ho & Salimans, 2021) is used, let $r_\tau^{\mathrm{u}} = f_\theta(s_\tau, \varnothing, t_\tau)$ and $r_\tau^{\mathrm{c}} = f_\theta(s_\tau, p, t_\tau)$ denote the unconditional and text-conditioned predictions. CFG forms

$$r_\tau^{\mathrm{cfg}} = r_\tau^{\mathrm{u}} + w\left(r_\tau^{\mathrm{c}} - r_\tau^{\mathrm{u}}\right), \tag{1}$$

where $w$ is the guidance scale. The sampler then advances the generative state according to

$$s_{\tau-1} = \Phi_\tau\left(s_\tau, r_\tau^{\mathrm{cfg}}\right), \tag{2}$$

where $\Phi_\tau$ denotes the model-specific sampling transition. Dependence on solver history, stochastic state, and other auxiliary variables is suppressed for clarity.

For continuous diffusion and flow models (Rombach et al., 2022; Lipman et al., 2023), $s_\tau = z_\tau \in \mathbb{R}^{C_z\times T_z\times H_z\times W_z}$, where $z_\tau$ is a real-valued spatiotemporal latent, typically spatially and possibly temporally compressed relative to $x$. The prediction $r_\tau$ is likewise real-valued and may represent noise, a score, or a velocity field; $\Phi_\tau$ is then the corresponding numerical solver step. For discrete diffusion and flow (Austin et al., 2021; Campbell et al., 2022; Lou et al., 2024; Gat et al., 2024), $s_\tau = y_\tau \in \{1, \dots, K\}^{T_y\times H_y\times W_y}$, where $K$ is the discrete vocabulary size. For URSA, $f_\theta$ predicts categorical logits $r_\tau = \ell_\tau \in \mathbb{R}^{T_y\times H_y\times W_y\times K}$, and $\Phi_\tau$ performs the categorical refinement.

### 3.2 SPATIOTEMPORAL ATTENTION

Video Transformers process spatiotemporal tokens $u_i \in \mathbb{R}^d$. Let $f(i)$ denote token $i$'s temporal position within the video, distinct from the inference step $\tau$.

For one self-attention head of dimension $d_h$, let $Q_i = u_i W_Q$, $K_j = u_j W_K$, and $V_j = u_j W_V$. Then

$$A_{ij} = \mathrm{softmax}_j \left( \frac{Q_i K_j^\top}{\sqrt{d_h}} \right), \qquad Y_i = \sum_j A_{ij} V_j, \tag{3}$$

where $i$ is the receiving token, $j$ is a source token, $A_{ij}$ determines how strongly source $j$ contributes, and $V_j$ is the feature content it provides. When $f(i) \neq f(j)$, self-attention transmits information between different temporal positions. Head and layer indices are omitted for clarity.

Let $h_1, \ldots, h_L$ denote the text-token features obtained by encoding prompt $p$. A controlled concept may span one or several tokenizer positions. Depending on the architecture, these text features condition video features through dedicated cross-attention or through joint text–video attention.

### 3.3 TEMPORAL CONTROL TASK

Given a text prompt $p$, we divide the video timeline into $T_c$ temporal intervals, called control bins, and assign a requested state to each controlled concept within each interval. For concepts $\mathcal{U} = \{q_1, \ldots, q_M\}$, these states form

$$\mathbf{C} = (c_{m,f}) \in \{0,1\}^{M \times T_c}, \tag{4}$$

where $m$ indexes concepts and $f$ indexes control bins. For an object, $c_{m,f} = 1$ requests presence in bin $f$, while $c_{m,f} = 0$ requests absence. For an action, these states indicate whether bin $f$ falls inside or outside the requested peak-motion interval; the action need not be absent outside that interval. The mapping $f(i)$ associates video token $i$ with a control bin according to its temporal position, giving it the requested state $c_{m,f(i)}$ for concept $q_m$.

Our experiments consider one time-varying target, whose schedule can specify onset, bounded presence, exit, or return. In the two-object setting, a second object remains present: $c_{S,f} = 1$, while the target follows $c_{T,f}$. Thus, the concept index distinguishes the timed target from persistent roles; the evaluated setting does not require multiple independently varying schedules.

## 4 CHRONOGUIDE

CG uses the control program $\mathbf{C}$ to enforce a concept's temporal schedule with a frozen generator. Figure 2 summarizes three stages: prepare inputs and text permissions, compute state-conditioned values, and exchange values across temporal states.

### 4.1 STATE-CONDITIONED BRANCHES AND CROSS-STATE ATTENTION

**Stage 1: Prepare inputs and text permissions.** Let $q_m$ be the time-varying target with schedule $c_{m,f}$; the two-object extension below handles persistent objects. At step $\tau$, the host embeds video state $s_\tau$ into tokens $U_\tau = (u_{\tau,1}, \ldots, u_{\tau,N_v})$ and prompt $p$ into text features $H_p = (h_1, \ldots, h_L)$. Token $i$ receives control state $\sigma_i = c_{m,f(i)}$.

We bind prompt spans to text-token sets (Appendix N): $J_m^{\mathrm{obj}}$ contains the target and its description; $J_m^{\mathrm{evt}}$ contains event spans excluded under event-aware absence masking; and $J_m^{\mathrm{exc}}$ contains phrases excluded under presence (labeled "cues" in the figure). The text permission masks are

$$\begin{aligned} \mu_{m,k}^{(0)} &= \mathbf{1}\left[k \notin J_m^{\mathrm{obj}} \cup J_m^{\mathrm{evt}}\right], \\ \mu_{m,k}^{(1)} &= \mathbf{1}\left[k \notin J_m^{\mathrm{exc}}\right]. \end{aligned} \tag{5}$$

These binary masks permit (1) or block (0) attention to text position $k$, without modifying $H_p$. Object-only absence uses $J_m^{\mathrm{evt}} = \varnothing$. Presence exclusions are specified independently and may include event phrases; with event-aware absence, $J_m^{\mathrm{exc}}$ and $J_m^{\mathrm{evt}}$ need not be disjoint.

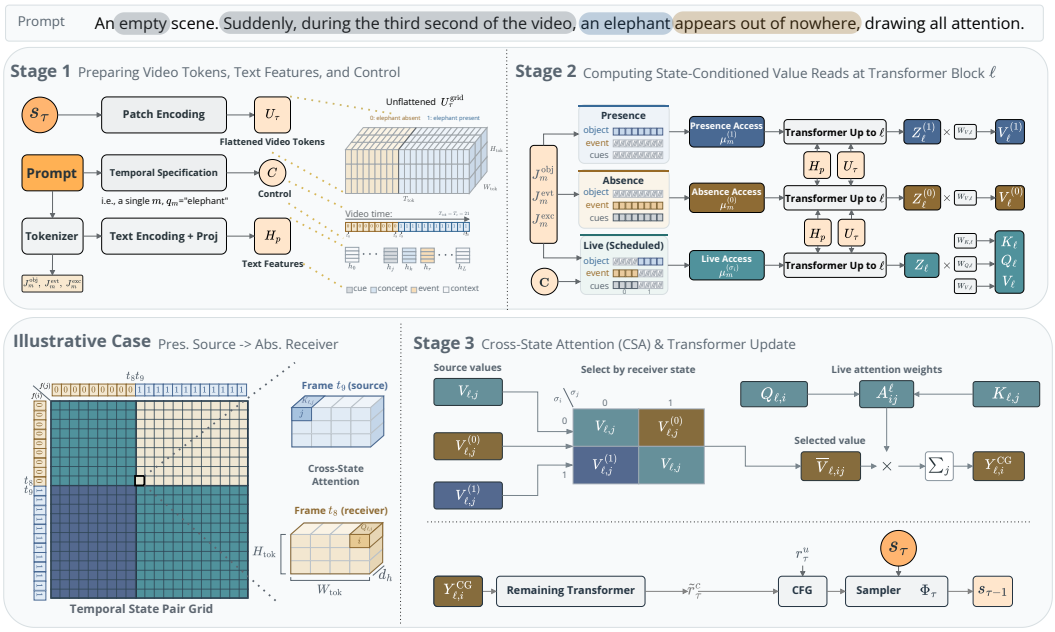



Figure 2: **ChronoGuide (CG) overview on Wan.** CG adds temporal control to frozen Wan through text permissions and value routing. **Stage 1:** Wan encodes video tokens $U_\tau$ and text features $H_p$; CG binds control **C** to video positions and prompt spans. **Stage 2:** Shared frozen branches compute live, absence-conditioned, and presence-conditioned values under different text permissions. **Stage 3:** CG retains live attention weights but replaces cross-state values according to the receiver's requested state; same-state values remain live. Only the live branch advances through the remaining Transformer, CFG, and sampler. Filled/hatched cells permit/block text access; the example uses object-only absence masking. Routing rows/columns denote receivers/sources.

**Stage 2: Compute state-conditioned values.** Three branches share $U_\tau$, $H_p$, the host timestep $t_\tau$, and frozen weights. In the live branch, query $i$ uses $\mu_m^{(\sigma_i)}$; the absence and presence branches use $\mu_m^{(0)}$ and $\mu_m^{(1)}$, respectively, at every video position.

At block $\ell$, let $Z_\ell$, $Z_\ell^{(0)}$, and $Z_\ell^{(1)}$ be the features entering self-attention. The shared value projection produces $V_\ell = \mathrm{Proj}_{V,\ell}(Z_\ell)$ and $V_\ell^{(b)} = \mathrm{Proj}_{V,\ell}(Z_\ell^{(b)})$, $b \in \{0,1\}$. Each source $j$ supplies a live value $V_{\ell,j}$ and two state-conditioned alternatives. The live branch also supplies $Q_\ell$ and $K_\ell$. Auxiliary branches compute their own internal attention without cross-state exchange and supply values rather than separate sampled videos.

**Stage 3: Exchange values and update the video state.** For one head at an intervened block, let $A_{ij}^\ell$ be the live attention weight from source $j$ to receiver $i$. CG preserves these weights and exchanges only cross-state video values. Let $\mathcal{V}$ denote video-token positions, with receiver $i \in \mathcal{V}$.

$$\overline{V}_{\ell,ij} = \begin{cases} V_{\ell,j}^{(\sigma_i)}, & j \in \mathcal{V},\ \sigma_i \neq \sigma_j, \\ V_{\ell,j}, & \text{otherwise.} \end{cases}$$
$$Y_{\ell,i}^{\mathrm{CG}} = \sum_j A_{ij}^\ell \overline{V}_{\ell,ij}. \tag{6}$$

Thus, an absence receiver reading a presence source uses $V_{\ell,j}^{(0)}$, and the reverse direction uses $V_{\ell,j}^{(1)}$. In joint text–video attention, text and other context sources retain their live values and remain in the native normalization. The modified output passes through the native output projection and subsequent layers, whose attention weights may change in response. Exchange repeats at the selected blocks; only the live branch supplies the conditional prediction to CFG and the sampler. Appearance,

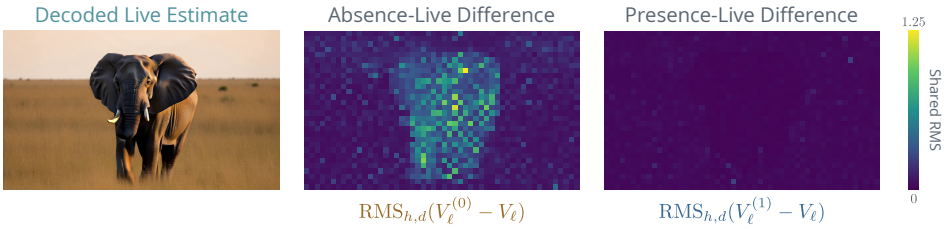



Figure 3: **State-conditioned values reflect object presence.** Left: decoded intermediate clean-state estimate at requested-present position $t_{10}$. Middle/right: per-token RMS differences $V_\ell^{(0)} - V_\ell$ and $V_\ell^{(1)} - V_\ell$ over heads and within-head channels, on a shared scale. Presence values closely match live values, while absence values differ along the elephant's silhouette, supporting state-conditioned value selection for temporal presence control. Object-only absence masking; update 30, block 15.

disappearance, and return differ only in their schedules. Figure 3 visualizes these state-conditioned value differences.

**Two-object composition.** Two-object scenes require maintaining $q_S$ while entrant $q_T$ follows $c_{T,f}$. During the first half of sampling, we supplement temporal routing with spatial text permissions, regional retrieval, and object-text amplification. This extension is used only for two objects. Preservation of live attention weights and same-state values refers to the core routing operator; the extension additionally adjusts composition (Appendix K).

### 4.2 Action Timing

Actor identity remains conditioned throughout. Action tokens $J_m^{\mathrm{act}}$ are masked outside the requested peak-motion interval ($\sigma_i = 0$) and retained and amplified inside it ($\sigma_i = 1$).

Let $\alpha_{ij}$ be the text-attention weight after applying branch-specific permissions and softmax normalization. In the live branch, we use

$$\widetilde{Z}_i^{\mathrm{text}} = \sum_j \alpha_{ij} V_j^{\mathrm{text}} + \sigma_i(\gamma - 1) \sum_{j \in J_m^{\mathrm{act}}} \alpha_{ij} \left(V_j^{\mathrm{text}} - b_V\right), \tag{7}$$

where $b_V$ is the value-projection bias and $\gamma \geq 1$ the action gain (Sec. 5). Auxiliary branch $b$ replaces $\sigma_i$ with $b$, applying the gain throughout the state-1 branch and never in the state-0 branch.

### 4.3 Optional Guidance Calibration

The core operator retains native CFG and sampling. For continuous models, we optionally rescale spatial bands of $\Delta_\tau = \widetilde{r}_\tau^{\mathrm{c}} - r_\tau^{\mathrm{u}}$, the live conditional prediction minus the unconditional prediction. Let $\mathcal{P}_{\mathrm{lo}}$ reconstruct the two-level spatial Haar coarse component and $b_f = \max_m c_{m,f}$. We use

$$r_\tau^{\mathrm{CG}}(f) = r_\tau^{\mathrm{u}}(f) + g_{\mathrm{lo}}^{(b_f)} \mathcal{P}_{\mathrm{lo}} \Delta_\tau(f) + g_{\mathrm{hi}} \big(I - \mathcal{P}_{\mathrm{lo}}\big) \Delta_\tau(f). \tag{8}$$

Here, $g_{\mathrm{lo}}^{(b_f)}$ and $g_{\mathrm{hi}}$ scale the coarse and detail components; $b_f$ includes persistent roles. We ablate this calibration separately. Sampling uses the native transition $\Phi_\tau$; discrete models retain native guidance and refinement.

## 5 Experiments

We evaluate CG on Wan 2.1-S (1.3B), CogVideoX-S (2B), and discrete URSA (1.7B). TempoControl (TC) (Schiber et al., 2026) provides 80 one-object (ONE), 82 two-object (TWO), and 100 movement prompts. Our mixed benchmark (MIX) covers all 80 COCO classes with TC-style prompts, manually chosen event phrases, and random intervals: 27 entry–exit, 27 exit, and 26 exit–return schedules. Published TC results are used where available; mixed-lifecycle and native URSA baselines are evaluated locally.

Table 1: TC benchmarks with CG (ours). Baselines are quoted from TC except local URSA. Metrics ↑; temporal scores in %. Bold: best per host; gains precede rounding. CG uses P for ONE and F for TWO; baseline scores are quoted as published. F-scored ONE: Appendix C.1.

| One-object method | Params / eval. | TA | Pres | Abs | MUSIQ |
|---|---|---|---|---|---|
| LTX-2 | 19B / P | 74.23 | 62.12 | 85.23 | 42.79 |
| CogX1.5-L | 5B / P | 61.43 | 75.50 | 48.64 | 65.36 |
| Wan2.1-L | 14B / P | 64.23 | 74.00 | 55.34 | 66.89 |
| Wan2.2 | 5B / P | 78.02 | 68.00 | 87.42 | 48.13 |
| CogX-S (native) | 2B / P | 58.15 | 39.88 | 74.77 | 45.16 |
| CogX-S + TC | 2B / P | 63.75 | 47.62 | 78.41 | 46.67 |
| CogX-S + CG (ours) | 2B / P | **75.06** | **69.50** | **80.63** | **49.53** |
| Wan2.1-S (native) | 1.3B / P | 63.94 | 60.50 | 67.38 | 53.76 |
| Wan2.1-S + TC | 1.3B / P | 83.56 | 79.75 | 87.38 | 56.92 |
| Wan2.1-S + CG (ours) | 1.3B / P | **90.06** | **83.88** | **96.25** | **57.12** |
| URSA (native) | 1.7B / P | 65.75 | 55.88 | 75.63 | 35.72 |
| URSA + CG (ours) | 1.7B / P | **73.69** | **63.63** | **83.75** | **37.90** |

| | Two-object | | | | Movement | |
|---|---|---|---|---|---|---|
| Method | TA | Pres | Abs | MUSIQ | TA | MUSIQ |
| LTX-2 | 39.31 | 16.08 | 64.88 | 58.62 | 28.00 | 51.61 |
| CogX1.5-L | 24.68 | 4.88 | 46.46 | 63.78 | 21.00 | 61.39 |
| Wan2.1-L | 40.07 | 27.27 | 54.15 | 69.01 | 18.00 | 61.31 |
| Wan2.2 | 47.05 | 36.59 | 56.86 | 67.83 | 23.00 | 60.61 |
| Wan2.1-S (native) | 37.50 | 29.15 | 45.85 | 68.56 | 19.00 | 60.46 |
| Wan2.1-S + TC | 53.17 | 49.02 | 57.32 | 70.82 | 54.00 | 63.24 |
| Wan2.1-S + CG (ours) | **67.83** | **60.42** | **75.98** | **70.89** | **61.00** | **66.97** |

We use B200 GPUs and seed 42, following TC, with frozen weights and native samplers. Wan uses 50 updates, Haar CFG 6/3/6, and action gain $\gamma = 1.5$ for movement. CogVideoX uses development-selected CFG 3.5; the default CFG 6 improves timing but lowers MUSIQ (Appendix J). Two-object and mixed-lifecycle details appear in Appendices K–L.

YOLOv10 (Wang et al., 2024a) uses 640-pixel inputs and threshold 0.25. TA averages video-level agreement with the requested schedule; Pres/Abs pool correct requested-present/absent decisions. ONE uses TC's released 20-frame evaluator (P); TWO follows its released evaluator with all 21 control entries (F), also used for MIX. MUSIQ (Ke et al., 2021) measures image quality, and movement TA follows TC's peak-flow-second metric.

For complementary human evaluation, 30 participants compare native Wan, TC, and CG on 18 randomly selected one-object, two-object, and mixed prompts (6 per task; 54 videos). Participants judge timing and visual quality with timestamped schedule aids and may select ties; participant-level favorites are their most-selected methods (Appendix A).

**Benchmark results.** CG improves all four one-object metrics over same-host TC on Wan and CogVideoX (+6.50/+11.31 TA), and over native URSA (+7.94 TA), supporting transfer across continuous and discrete hosts (Table 1). For two objects, composition support raises the event-aware core's TA/Pres from 57.26/42.24 to 67.83/60.42, exceeding TC's published 53.17/49.02 (Appendix K). Movement gains 7.00 TA and 3.73 MUSIQ. On mixed lifecycles, CG gains 16.07 TA, 3.67 Pres, 30.31 Abs, and 1.69 MUSIQ (Table 2). Thus the large absence gain accompanies improved required presence, rather than suppressing the object throughout the video.

CG is the most frequent participant favorite for both timing and visual quality across the human study. It is the unique timing favorite for 28/30 participants (93.3%) and the unique quality favorite for 15/30 (50.0%), versus 3.3% and 26.7% for TC. Including shared favorites, CG reaches 96.7% and 60.0%, respectively (Appendix A).

Table 2: **Mixed lifecycles.** 80 prompts, YOLOv10, full-timeline scoring (F); TC uses $\ell = 5$.

| Method | TA ↑ | Pres ↑ | Abs ↑ | MUSIQ ↑ |
|---|---|---|---|---|
| Wan2.1-S (native) | 57.68 | 73.50 | 39.51 | 60.21 |
| Wan2.1-S + TC | 66.37 | 74.61 | 56.91 | 59.71 |
| Wan2.1-S + CG (ours) | **82.44** | **78.29** | **87.21** | **61.39** |

Table 3: **Component effects.** Exchange: on−off (Haar fixed); Haar: bandwise−CFG 6 (exchange on); Event: identity+event−identity (exchange/Haar on). MIX uses event masking unless noted. Blue/amber: gain/loss; temporal metrics in pp, MUSIQ in points. Details: Appendix D.

| | Seed 42 | | | | Seed 43 | | | |
|---|---|---|---|---|---|---|---|---|
| Component | ΔTA | ΔPres | ΔAbs | ΔMUSIQ | ΔTA | ΔPres | ΔAbs | ΔMUSIQ |
| Exchange: ONE | +6.50 | +5.11 | +7.59 | -0.63 | +3.75 | +9.66 | -0.89 | +7.38 |
| Exchange: MIX, identity | +28.57 | +13.41 | +39.83 | -2.40 | +15.48 | +5.03 | +23.24 | -1.01 |
| Exchange: MIX, +event | +27.14 | +3.91 | +44.40 | -1.62 | +21.67 | +10.61 | +29.88 | +2.79 |
| Haar CFG: ONE | +3.00 | +7.95 | -0.89 | +1.17 | -1.75 | +6.82 | -8.48 | +7.71 |
| Haar CFG: MIX | -0.48 | +15.64 | -12.45 | +7.36 | -0.95 | -5.03 | +2.07 | +5.44 |
| Event mask: MIX | -2.62 | +1.68 | -5.81 | +2.50 | -0.95 | +2.23 | -3.32 | +4.79 |

**Seed and detector robustness.** Across seeds 42–46, YOLO TA is $87.31 \pm 1.89$ (SD; 95% CI $[84.97, 89.66]$), exceeding published TC's 83.56 at every random seed. All four metric means are higher; TA and Abs exceed the fixed references after Holm correction ($p = 0.034$ each; Appendix B). OWLv2 (Minderer et al., 2023) scores the same videos at $91.80 \pm 2.55$ TA and $92.25 \pm 3.89$ Pres (Appendix C).

### 5.1 Model analysis and cost

**Component effects.** Four ONE and six MIX configurations use twenty matched prompts and seeds 42/43 (400 videos). Exchange improves TA under Haar (Table 3) and scalar CFG (ONE: +2.00/+6.25; MIX: +14.28/+19.76), supporting a benefit beyond text masking without requiring Haar. Haar raises MUSIQ in every task/seed pair; event masking increases MIX Pres/MUSIQ while reducing TA/Abs (Appendix D). In the 96-video action ablation, exchange adds 16.67 points in peak-motion TA at both seeds with unit text gain; its effect varies at $\gamma = 1.5$ (Appendix I).

**Cost and scalability.** On one B200, CG takes 213.34/141.64/98.63 s for C50/C25/C10. C25/C10 save 33.6%/53.8% relative to C50 while retaining all 50 solver updates. TC reports ∼307 s on B200 (same GPU model) (Schiber et al., 2026); timing sources and measurement scopes appear in Appendix G.2. CG matches native Wan's batch-one allocated memory: 21.94/10.83 GiB resident/offloaded, versus 114.91/104.32 GiB for local TC—81%/90% less. This enables inference on lower-cost hardware and batched generation: CG runs on 24 GB RTX 3090/4090 GPUs and supports batched generation on B200, where a measured batch of ten uses only 33.55 GiB with offloading (Appendix G).

**Why receiver-matched routing?** Four controls add 280 videos paired with 120 CG baselines (Table 4). CG gives the strongest temporal control: every ONE/MIX alternative lowers TA at both seeds. Blocking loses 35.75–53.07 points of required presence; global replacement gains 1.53/0.98 MUSIQ but loses 8.88/5.95 TA. Thus, we retain selective routing in CG, whose headline results improve both TA and MUSIQ over same-host TC. For TWO, exchange raises mean TA/Abs by 6.43/12.00 with composition controls fixed; mean MUSIQ also improves. Paired intervals include zero (Appendix D.6).

**Guidance duration.** On twenty prompts/task (seed 42), CG runs for 10/25/40/50 updates, followed by native conditioning or masks/Haar without exchange (Figure 4). C10 with native continuation improves ONE TA/MUSIQ by 1.75/0.51 over C50, preserving Abs. MIX favors C50 (+3.57

Table 4: **Ablating the routing rule.** ONE/MIX: twenty prompts, seeds 42/43 averaged. TWO: forty prompts, seed 42. Development panels; metrics ↑. Shaded rows: full CG reference. Paired intervals: Appendix D.

| | ONE | | MIX | |
|---|---|---|---|---|
| Routing variant | TA | MUSIQ | TA | MUSIQ |
| **CG (ours)** | 83.50 | 63.13 | 78.57 | 64.06 |
| Wrong-state donors | 77.12 | 60.71 | 56.31 | 64.84 |
| Cross-state blocking | 68.88 | 57.94 | 67.50 | 56.71 |
| Global replacement | 74.62 | 64.66 | 72.62 | 65.04 |
| TWO | Joint TA | Abs | Incumbent Pres | MUSIQ |
| **CG (ours)** | 61.79 | 72.25 | 81.19 | 71.50 |
| Cross-state exchange off | 55.36 | 60.25 | 80.71 | 70.94 |

TA, +3.73 Abs); masks/Haar alone do not consistently recover this gain. This favors short guidance for simple appearances and continued exchange for exits/returns; headline runs use C50.

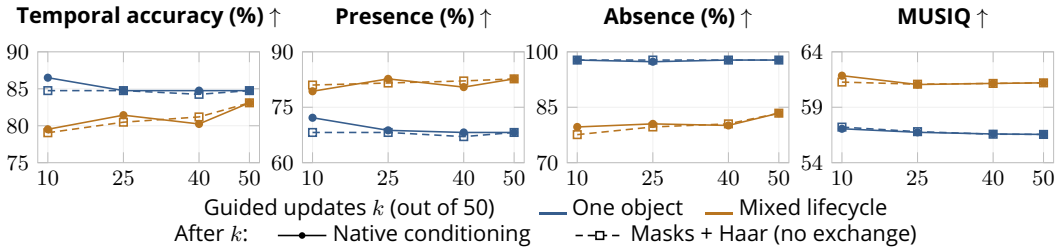



Figure 4: **Guidance duration.** After $k$ updates: native conditioning (circles) or masks/Haar without exchange (squares). All runs use 50 updates. ONE: P; MIX: F; ↑: better. Appendix E.

**Transitions and dynamics.** Across 80 mixed exits, CG leads in every post-exit bin, reaching 77.5% correct absence at 0.75 s versus TC's 50.0% and native Wan's 31.25% (Figure 5a). Relative RMS correction falls from 6.2–6.8% early to about 1.2% late across sixteen videos (Figure 5b), motivating short guidance; the MIX cutoff results support continued exchange for complex schedules.

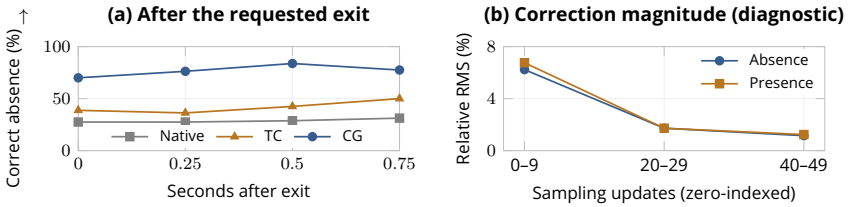



Figure 5: **Transitions and denoising dynamics.** (a) Correct absence after mixed exits (↑ better). (b) Relative RMS exchange correction: 16 videos, seeds 42/43. Details: Appendices F, H.

## 6 Conclusion

ChronoGuide identifies cross-state feature propagation in spatiotemporal self-attention as a source of temporal-control errors and addresses it through training-free cross-state value routing. It improves every headline metric over same-host TempoControl on one-object, two-object, movement, and mixed-lifecycle benchmarks, at a fraction of the peak memory, and transfers to continuous and discrete hosts. Reducing the cost of the auxiliary branches and extending control to several independently timed concepts are natural next steps.

### AI USE STATEMENT

Generative AI tools were used to assist with writing and improving clarity, and to assist with the development and refactoring of code and experimental scripts. The scientific ideas, methodology, experimental design, results, and conclusions were developed and validated by the authors. All parts where AI tools were involved were reviewed, edited, and verified by the authors.

### REPRODUCIBILITY STATEMENT

The supplementary material includes CG implementations for Wan, URSA, and CogVideoX, together with benchmark prompts, temporal schedules, evaluation code, configuration files, pinned dependencies, and container recipes. It also documents all inference settings and selected seeds, and provides a comparison website with videos and frame strips. We validate the implementation with automated tests and representative comparisons against the research code.

### ETHICS STATEMENT

The human evaluation involved 30 participants recruited among university students and colleagues, without compensation. Participants were informed that the study supported a research paper and that their evaluations would be reported. No participant names or other identifying information are disclosed in the paper.

## REFERENCES

Donghoon Ahn, Hyoungwon Cho, Jaewon Min, Wooseok Jang, Jungwoo Kim, SeonHwa Kim, Hyun Hee Park, Kyong Hwan Jin, and Seungryong Kim. Self-rectifying diffusion sampling with perturbed-attention guidance. In *European Conference on Computer Vision*, pp. 1–17. Springer, 2024.

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. Structured denoising diffusion models in discrete state-spaces. *Advances in neural information processing systems*, 34:17981–17993, 2021.

Sherwin Bahmani, Ivan Skorokhodov, Aliaksandr Siarohin, Willi Menapace, Guocheng Qian, Michael Vasilkovsky, Hsin-Ying Lee, Chaoyang Wang, Jiaxu Zou, Andrea Tagliasacchi, David B. Lindell, and Sergey Tulyakov. VD3D: Taming large video diffusion transformers for 3D camera control. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=0n4bS0R5MM.

Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 843–852, June 2023.

Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. MultiDiffusion: Fusing diffusion paths for controlled image generation. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 1737–1752. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/bar-tal23a.html.

Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, et al. Lumiere: A space-time diffusion model for video generation. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets, 2023a. URL https://arxiv.org/abs/2311.15127.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align Your Latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22563–22575, June 2023b.

Manuel Brack, Felix Friedrich, Dominik Hintersdorf, Lukas Struppek, Patrick Schramowski, and Kristian Kersting. SEGA: Instructing text-to-image models using semantic guidance. *Advances in Neural Information Processing Systems*, 36:25365–25389, 2023.

Minghong Cai, Xiaodong Cun, Xiaoyu Li, Wenze Liu, Zhaoyang Zhang, Yong Zhang, Ying Shan, and Xiangyu Yue. DiTCtrl: Exploring attention control in multi-modal diffusion transformer for tuning-free multi-prompt longer video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7763–7772, 2025.

Andrew Campbell, Joe Benton, Valentin De Bortoli, Thomas Rainforth, George Deligiannidis, and Arnaud Doucet. A continuous time framework for discrete denoising models. *Advances in Neural Information Processing Systems*, 35:28266–28279, 2022.

Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. MasaCtrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 22560–22570, 2023.

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-Excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Trans. Graph.*, 42 (4), July 2023. ISSN 0730-0301. doi: 10.1145/3592116. URL https://doi.org/10.1145/3592116.

Gordon Chen, Ziqi Huang, and Ziwei Liu. Prompt relay: Inference-time temporal control for multi-event video generation, 2026. URL https://arxiv.org/abs/2604.10030.

Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. VideoCrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7310–7320, June 2024.

Hyungjin Chung, Jeongsol Kim, Michael Thompson Mccann, Marc Louis Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=OnD9zGAGT0k.

Omer Dahary, Or Patashnik, Kfir Aberman, and Daniel Cohen-Or. Be Yourself: Bounded attention for multi-subject text-to-image generation. In *European Conference on Computer Vision*, pp. 432–448. Springer, 2024.

Haoge Deng, Ting Pan, Fan Zhang, Yang Liu, Zhuoyan Luo, Yufeng Cui, Wenxuan Wang, Chunhua Shen, Shiguang Shan, Zhaoxiang Zhang, and Xinlong Wang. Uniform discrete diffusion with metric path for video generation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=GFU5yCbILk.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Rotem Ezra, Hedi Zisling, Nimrod Berman, Ilan Naiman, Alexey Gorkor, Liran Nochumsohn, Eliya Nachmani, and Omri Azencot. FreeSliders: Training-free, modality-agnostic concept sliders for fine-grained diffusion control in images, audio, and video, 2025. URL https://arxiv.org/abs/2511.00103.

Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Proceedings of the 13th Scandinavian Conference on Image Analysis*, LNCS 2749, pp. 363–370, Gothenburg, Sweden, June-July 2003.

Weixi Feng, Chao Liu, Sifei Liu, William Yang Wang, Arash Vahdat, and Weili Nie. BlobGEN-Vid: Compositional text-to-video generation with blob video representations. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12989–12998, 2025.

Itai Gat, Tal Remez, Neta Shaul, Felix Kreuk, Ricky TQ Chen, Gabriel Synnaeve, Yossi Adi, and Yaron Lipman. Discrete flow matching. *Advances in Neural Information Processing Systems*, 37: 133345–133385, 2024.

Daniel Geng, Charles Herrmann, Junhwa Hur, Forrester Cole, Serena Zhang, Tobias Pfaff, Tatiana Lopez-Guevara, Yusuf Aytar, Michael Rubinstein, Chen Sun, et al. Motion Prompting: Controlling video generation with motion trajectories. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 1–12, 2025.

Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. TokenFlow: Consistent diffusion features for consistent video editing. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=lKK50q2MtV.

Sharath Girish, Viacheslav Ivanov, Tsai-Shien Chen, Hao Chen, Aliaksandr Siarohin, and Sergey Tulyakov. AlcheMinT: Fine-grained temporal control for multi-reference consistent video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23227–23237, 2026.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. AnimateDiff: Animate your personalized text-to-image diffusion models without specific tuning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Fx2SbBgcte.

Agrim Gupta, Lijun Yu, Kihyuk Sohn, Xiuye Gu, Meera Hahn, Fei-Fei Li, Irfan Essa, Lu Jiang, and José Lezama. Photorealistic video generation with diffusion models. In *European Conference on Computer Vision*, pp. 393–411. Springer, 2024.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, Eitan Richardson, Guy Shiran, Itay Chachy, Jonathan Chetboun, Michael Finkelson, Michael Kupchick, Nir Zabari, Nitzan Guetta, Noa Kotler, Ofir Bibi, Ori Gordon, Poriya Panet, Roi Benita, Shahar Armon, Victor Kulikov, Yaron Inger, Yonatan Shiftan, Zeev Melumian, and Zeev Farbman. LTX-2: Efficient joint audio-visual foundation model, 2026. URL https://arxiv.org/abs/2601.03233.

Hui Han, Siyuan Li, Jiaqi Chen, Yiwen Yuan, Yuling Wu, Yufan Deng, Chak Tou Leong, Hanwen Du, Junchen Fu, Youhua Li, et al. Video-Bench: Human-aligned video generation benchmark. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18858–18868, 2025.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. CameraCtrl: Enabling camera control for video diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Z4evOUYrk7.

Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, et al. VideoScore: Building automatic metrics to simulate fine-grained human feedback for video generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2105–2123, 2024a.

Yutong He, Naoki Murata, Chieh-Hsin Lai, Yuhta Takida, Toshimitsu Uesaka, Dongjun Kim, Wei-Hsiang Liao, Yuki Mitsufuji, J Zico Kolter, Ruslan Salakhutdinov, and Stefano Ermon. Manifold preserving guided diffusion. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=o3BxOLoxm1.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross-attention control. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=_CDixzkzeyb.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. URL https://openreview.net/forum?id=qw8AKxfYbI.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen Video: High definition video generation with diffusion models, 2022a. URL https://arxiv.org/abs/2210.02303.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David Fleet. Video diffusion models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 8633–8646. Curran Associates, Inc., 2022b. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/39235c56aef13fb05a6adc95eb9d8d66-Paper-Conference.pdf.

Zhaoyang Huang, Xiaoyu Shi, Chao Zhang, Qiang Wang, Ka Chun Cheung, Hongwei Qin, Jifeng Dai, and Hongsheng Li. FlowFormer: A transformer architecture for optical flow. In *European conference on computer vision*, pp. 668–685. Springer, 2022.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024.

Junha Hyung, Kinam Kim, Susung Hong, Min-Jung Kim, and Jaegul Choo. Spatiotemporal skip guidance for enhanced video diffusion sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11006–11015, June 2025.

Yash Jain, Anshul Nasery, Vibhav Vineet, and Harkirat Behl. PEEKABOO: Interactive video generation via masked-diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8079–8088, 2024.

Sangwon Jang, Taekyung Ki, Jaehyeong Jo, Jaehong Yoon, Soo Ye Kim, Zhe Lin, and Sung Ju Hwang. Frame Guidance: Training-free guidance for frame-level control in video diffusion models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=y39XbEp1vK.

Longbin Ji, Lei Zhong, Pengfei Wei, and Changjian Li. PoseTraj: Pose-aware trajectory control in video diffusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 22776–22785, 2025.

Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong MU, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=66NzcRQuOq.

Taewon Kang and Matthias Zwicker. Tpd: Temporal prior decoupling for text-to-video diffusion models, 2026. URL https://arxiv.org/abs/2607.26706.

Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. *Advances in Neural Information Processing Systems*, 37:52996–53021, 2024.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5148–5157, October 2021.

Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2Video-Zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15954–15964, 2023.

Jaemin Kim, Bryan Sangwoo Kim, and Jong Chul Ye. Free2Guide: Training-free text-to-video alignment using image LVLM. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17920–17929, October 2025.

Dan Kondratyuk, Lijun Yu, Xiuye Gu, Jose Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, Krishna Somandepalli, Hassan Akbari, Yair Alon, Yong Cheng, Joshua V. Dillon, Agrim Gupta, Meera Hahn, Anja Hauth, David Hendon, Alonso Martinez, David Minnen, Mikhail Sirotenko, Kihyuk Sohn, Xuan Yang, Hartwig Adam, Ming-Hsuan Yang, Irfan Essa, Huisheng Wang, David A Ross, Bryan Seybold, and Lu Jiang. VideoPoet: A large language model for zero-shot video generation. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 25105–25124. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/kondratyuk24a.html.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Junkun Yuan, Yanxin Long, Aladdin Wang, Andong Wang, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Hongmei Wang, Jacob Song, Jiawang Bai, Jianbing Wu, Jinbao Xue, Joey Wang, Kai Wang, Mengyang Liu, Pengyu Li, Shuai Li, Weiyan Wang, Wenqing Yu, Xinchi Deng, Yang Li, Yi Chen, Yutao Cui, Yuanbo Peng, Zhentao Yu, Zhiyu He, Zhiyong Xu, Zixiang Zhou, Zunnan Xu, Yangyu Tao, Qinglin Lu, Songtao Liu, Dax Zhou, Hongfa Wang, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, and Caesar Zhong. HunyuanVideo: A systematic framework for large video generative models, 2025. URL https://arxiv.org/abs/2412.03603.

Tuomas Kynkäänniemi, Miika Aittala, Tero Karras, Samuli Laine, Timo Aila, and Jaakko Lehtinen. Applying guidance in a limited interval improves sample and distribution quality in diffusion models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=nAIhvNy15T.

Dohun Lee, Bryan Sangwoo Kim, Geon Yeong Park, and Jong Chul Ye. VideoGuide: Improving video diffusion models without training through a teacher's guide. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2599–2608, June 2025.

Guojun Lei, Chi Wang, Rong Zhang, Yikai Wang, Hong Li, and Weiwei Xu. AnimateAnything: Consistent and controllable animation for video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27946–27956, June 2025.

Quanhao Li, Zhen Xing, Rui Wang, Hui Zhang, Qi Dai, and Zuxuan Wu. MagicMotion: Controllable video generation with dense-to-sparse trajectory guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12112–12123, 2025.

Long Lian, Baifeng Shi, Adam Yala, Trevor Darrell, and Boyi Li. LLM-grounded video diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=exKHibougU.

Mingxiang Liao, Hannan Lu, Xinyu Zhang, Fang Wan, Tianyu Wang, Yuzhong Zhao, Wangmeng Zuo, Qixiang Ye, and Jingdong Wang. Evaluation of text-to-video generation models: A dynamics perspective. *Advances in Neural Information Processing Systems*, 37:109790–109816, 2024.

Ruotong Liao, Guowen Huang, Qing Cheng, Guangyao Zhai, Lei Zhang, Xun Xiao, Thomas Seidl, Daniel Cremers, and Volker Tresp. TunerDiT: Training-free progressive steering of diffusion transformer for multi-event video generation. *arXiv preprint arXiv:2605.31590*, 2026.

Han Lin, Abhay Zala, Jaemin Cho, and Mohit Bansal. VideodirectorGPT: Consistent multi-scene video generation via LLM-guided planning. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=sKNIjS2brr.

Pengyang Ling, Jiazi Bu, Pan Zhang, Xiaoyi Dong, Yuhang Zang, Tong Wu, Huaian Chen, Jiaqi Wang, and Yi Jin. MotionClone: Training-free motion cloning for controllable video generation. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.),

*International Conference on Learning Representations*, volume 2025, pp. 75579–75601, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/bc82dbfbfa43232be85b8d9838f49c3e-Paper-Conference.pdf.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *European conference on computer vision*, pp. 423–439. Springer, 2022.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024a.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=XVjTT1nw5z.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. EvalCrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22139–22149, 2024b.

Yuanxin Liu, Rui Zhu, Shuhuai Ren, Jiacong Wang, Haoyuan Guo, Xu Sun, and Lu Jiang. UVE: Are MLLMs unified evaluators for AI-generated videos? *Advances in Neural Information Processing Systems*, 38, 2025.

Ling Lo, Kelvin CK Chan, Wen-Huang Cheng, and Ming-Hsuan Yang. From prompt to progression: Taming video diffusion models for seamless attribute transition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18651–18660, 2025.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32819–32848. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/lou24a.html.

Wan-Duo Kurt Ma, J. P. Lewis, and W. Bastiaan Kleijn. TrailBlazer: Trajectory control for diffusion-based video generation. In *SIGGRAPH Asia 2024 Conference Papers*, SA '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400711312. doi: 10.1145/3680528.3687652. URL https://doi.org/10.1145/3680528.3687652.

Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection. In *European conference on computer vision*, pp. 728–755. Springer, 2022.

Matthias Minderer, Alexey A. Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=mQPNcBWjGc.

Sicheng Mo, Fangzhou Mu, Kuan Heng Lin, Yanli Liu, Bochen Guan, Yin Li, and Bolei Zhou. FreeControl: Training-free spatial control of any text-to-image diffusion model with any condition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7465–7475, 2024.

Chong Mou, Xintao Wang, Jiechong Song, Ying Shan, and Jian Zhang. DragonDiffusion: Enabling drag-style manipulation on diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=OEL4FJMg1b.

Hunter Nisonoff, Junhao Xiong, Stephan Allenspach, and Jennifer Listgarten. Unlocking guidance for discrete state-space diffusion and flow models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=XsgHl54yO7.

Gyeongrok Oh, Jaehwan Jeong, Sieun Kim, Wonmin Byeon, Jinkyu Kim, Sungwoong Kim, and Sangpil Kim. Mevg: Multi-event video generation with text-to-video models. In *European Conference on computer vision*, pp. 401–418. Springer, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205, October 2023.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, David Yan, Dhruv Choudhary, Dingkang Wang, Geet Sethi, Guan Pang, Haoyu Ma, Ishan Misra, Ji Hou, Jialiang Wang, Kiran Jagadeesh, Kunpeng Li, Luxin Zhang, Mannat Singh, Mary Williamson, Matt Le, Matthew Yu, Mitesh Kumar Singh, Peizhao Zhang, Peter Vajda, Quentin Duval, Rohit Girdhar, Roshan Sumbaly, Sai Saketh Rambhatla, Sam Tsai, Samaneh Azadi, Samyak Datta, Sanyuan Chen, Sean Bell, Sharadh Ramaswamy, Shelly Sheynin, Siddharth Bhattacharya, Simran Motwani, Tao Xu, Tianhe Li, Tingbo Hou, Wei-Ning Hsu, Xi Yin, Xiaoliang Dai, Yaniv Taigman, Yaqiao Luo, Yen-Cheng Liu, Yi-Chiao Wu, Yue Zhao, Yuval Kirstain, Zecheng He, Zijian He, Albert Pumarola, Ali Thabet, Artsiom Sanakoyeu, Arun Mallya, Baishan Guo, Boris Araya, Breena Kerr, Carleigh Wood, Ce Liu, Cen Peng, Dimitry Vengertsev, Edgar Schonfeld, Elliot Blanchard, Felix Juefei-Xu, Fraylie Nord, Jeff Liang, John Hoffman, Jonas Kohler, Kaolin Fire, Karthik Sivakumar, Lawrence Chen, Licheng Yu, Luya Gao, Markos Georgopoulos, Rashel Moritz, Sara K. Sampson, Shikai Li, Simone Parmeggiani, Steve Fine, Tara Fowler, Vladan Petrovic, and Yuming Du. Movie gen: A cast of media foundation models, 2025. URL https://arxiv.org/abs/2410.13720.

Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. FateZero: Fusing attentions for zero-shot text-based video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15932–15942, 2023.

Tianhao Qi, Jianlong Yuan, Wanquan Feng, Shancheng Fang, Jiawei Liu, SiYu Zhou, Qian He, Hongtao Xie, and Yongdong Zhang. Mask2DiT: Dual mask-based diffusion transformer for multi-scene long video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18837–18846, 2025.

Haonan Qiu, Menghan Xia, Yong Zhang, Yingqing He, Xintao Wang, Ying Shan, and Ziwei Liu. FreeNoise: Tuning-free longer video diffusion via noise rescheduling. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ijoqFqSC7p.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. GEN3C: 3D-informed world-consistent video generation with precise camera control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6121–6132, 2025.

Kevin Rojas, Ye He, Chieh-Hsin Lai, Yuhta Takida, Yuki Mitsufuji, and Molei Tao. Improving classifier-free guidance in masked diffusion: Low-dim theoretical insights with high-dim impact. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=mMK9pvQJxf.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

Seyedmorteza Sadat, Jakob Buhmann, Derek Bradley, Otmar Hilliges, and Romann M. Weber. CADS: Unleashing the diversity of diffusion models through condition-annealed sampling. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=zMoNrajk2X.

Shira Schiber, Ofir Lindenbaum, and Idan Schwartz. TempoControl: Temporal attention guidance for text-to-video models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 36670–36679, June 2026.

Ariel Shaulov, Itay Hazan, Lior Wolf, and Hila Chefer. FlowMo: Variance-based flow guidance for coherent motion in video generation. *Advances in Neural Information Processing Systems*, 38: 82725–82754, 2025.

Xiaoyu Shi, Zhaoyang Huang, Fu-Yun Wang, Weikang Bian, Dasong Li, Yi Zhang, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, et al. Motion-I2V: Consistent and controllable image-to-video generation with explicit motion modeling. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11, 2024.

Zhilei Shu, Shangwen Zhu, Zihang Liang, Xiaofan Li, Qianyu Peng, Xinyu Cui, Bo Ye, Yiming Li, Fan Cheng, Jian Zhao, Yang Cao, Zheng-Jun Zha, and Ruili Feng. TIE: Time interval encoding for video generation over events, 2026. URL https://arxiv.org/abs/2605.10543.

Assaf Singer, Noam Rotstein, Amir Mann, Ron Kimmel, and Or Litany. Time-to-move: Training-free motion-controlled video generation via dual-clock denoising. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=OxO9OSYVw5.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-A-Video: Text-to-video generation without text-video data. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=nJfylDvgzlq.

Ivan Skorokhodov, Willi Menapace, Aliaksandr Siarohin, and Sergey Tulyakov. Hierarchical patch diffusion models for high-resolution video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7569–7579, 2024.

Jiaming Song, Qinsheng Zhang, Hongxu Yin, Morteza Mardani, Ming-Yu Liu, Jan Kautz, Yongxin Chen, and Arash Vahdat. Loss-guided diffusion models for plug-and-play controllable generation. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 32483–32498. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/song23k.html.

Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. T2V-CompBench: A comprehensive benchmark for compositional text-to-video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8406–8416, 2025.

Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pp. 402–419. Springer, 2020.

Yoad Tewel, Omri Kaduri, Rinon Gal, Yoni Kasten, Lior Wolf, Gal Chechik, and Yuval Atzmon. Training-free consistent text-to-image generation. *ACM Transactions on Graphics (TOG)*, 43(4): 1–18, 2024.

Ye Tian, Ling Yang, Haotian Yang, Yuan Gao, Yufan Deng, Jingmin Chen, Xintao Wang, Zhaochen Yu, Xin Tao, Pengfei Wan, et al. VideoTetris: Towards compositional text-to-video generation. *Advances in Neural Information Processing Systems*, 37:29489–29513, 2024.

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1921–1930, 2023.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. FVD: A new metric for video generation, 2019. URL https://openreview.net/forum?id=rylgEULtdN.

Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual descriptions. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=vOEXS39nOF.

Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/04025959b191f8f9de3f924f0940515f-Paper.pdf.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models, 2025. URL https://arxiv.org/abs/2503.20314.

Ao Wang, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, and Guiguang Ding. YOLOv10: Real-time end-to-end object detection. *Advances in neural information processing systems*, 37: 107984–108011, 2024a.

Jiawei Wang, Yuchen Zhang, Jiaxin Zou, Yan Zeng, Guoqiang Wei, Liping Yuan, and Hang Li. Boximator: Generating rich and controllable motions for video synthesis. In *Forty-first International Conference on Machine Learning*, 2024b. URL https://openreview.net/forum?id=7wgXuNOF0V.

Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. VideoComposer: Compositional video synthesis with motion controllability. *Advances in Neural Information Processing Systems*, 36:7594–7611, 2023.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. InternVideo2: Scaling foundation models for multimodal video understanding. In *European conference on computer vision*, pp. 396–416. Springer, 2024c.

Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. MotionCtrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, SIGGRAPH '24, New York, NY, USA, 2024d. Association for Computing Machinery. ISBN 9798400705250. doi: 10.1145/3641519.3657518. URL https://doi.org/10.1145/3641519.3657518.

Ziyi Wu, Aliaksandr Siarohin, Willi Menapace, Ivan Skorokhodov, Yuwei Fang, Varnith Chordia, Igor Gilitschenski, and Sergey Tulyakov. Mind the Time: Temporally-controlled multi-event video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23989–24000, June 2025.

Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, and Dacheng Tao. GMFlow: Learning optical flow via global matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8121–8130, 2022.

Qianxun Xu, Chenxi Song, Yujun Cai, and Chi Zhang. SwitchCraft: Training-free multi-event video generation with attention controls. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29136–29145, 2026.

Shiyuan Yang, Liang Hou, Haibin Huang, Chongyang Ma, Pengfei Wan, Di Zhang, Xiaodong Chen, and Jing Liao. Direct-a-Video: Customized video generation with user-directed camera movement and object motion. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–12, 2024.

Yuhang Yang, Ke Fan, Shangkun Sun, Hongxiang Li, Ailing Zeng, FeiLin Han, Wei Zhai, Wei Liu, Yang Cao, and Zheng-Jun Zha. VideoGen-Eval: Agent-based system for video generation evaluation, 2025a. URL https://arxiv.org/abs/2503.23452.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. CogVideoX: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, volume 2025, pp. 83048–83077, 2025b.

Haotian Ye, Haowei Lin, Jiaqi Han, Minkai Xu, Sheng Liu, Yitao Liang, Jianzhu Ma, James Zou, and Stefano Ermon. TFG: Unified training-free guidance for diffusion models. *Advances in Neural Information Processing Systems*, 37:22370–22417, 2024.

Jiwen Yu, Yinhuai Wang, Chen Zhao, Bernard Ghanem, and Jian Zhang. FreeDoM: Training-free energy-guided conditional diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23174–23184, 2023.

Hongyu Zhang, Yufan Deng, Zilin Pan, Peng-Tao Jiang, Bo Li, Qibin Hou, Zhen Dong, Zhiyang Dou, and Daquan Zhou. TS-Attn: Temporal-wise separable attention for multi-event video generation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=QixNhagZ9t.

Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, XIAOPENG ZHANG, Wangmeng Zuo, and Qi Tian. ControlVideo: Training-free controllable text-to-video generation. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=5a79AqFr0c.

Yang Zhang, Teoh Tze Tzun, Lim Wei Hern, and Kenji Kawaguchi. Enhancing semantic fidelity in text-to-image synthesis: Attention regulation in diffusion models. In *European Conference on Computer Vision*, pp. 70–86. Springer, 2024b.

Zhenghao Zhang, Junchao Liao, Menghao Li, Zuozhuo Dai, Bingxue Qiu, Siyu Zhu, Long Qin, and Weizhi Wang. Tora: Trajectory-oriented diffusion transformer for video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2063–2073, 2025.

Zangwei Zheng, Xiangyu Peng, Yuxuan Lou, Chenhui Shen, Tom Young, Xinying Guo, Binluo Wang, Hang Xu, Hongxin Liu, Mingyan Jiang, Wenjun Li, Yuhui Wang, Anbang Ye, Gang Ren, Qianran Ma, Wanying Liang, Xiang Lian, Xiwen Wu, Yuting Zhong, Zhuangyan Li, Chaoyu Gong, Guojun Lei, Leijun Cheng, Limin Zhang, Minghao Li, Ruijie Zhang, Silan Hu, Shijie Huang, Xiaokang Wang, Yuanheng Zhao, Yuqi Wang, Ziang Wei, and Yang You. Open-Sora 2.0: Training a commercial-level video generation model in $200k, 2026. URL https://arxiv.org/abs/2503.09642.

Xingyu Zhou, Qifan Li, Xiaobin Hu, Hai Chen, and Shuhang Gu. Guiding a diffusion transformer with the internal dynamics of itself. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11536–11545, June 2026.

Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation. *Advances in Neural Information Processing Systems*, 37:110315–110340, 2024.

## GUIDE TO THE SUPPLEMENTARY MATERIAL

The main paper contains the full benchmark comparison, human-study headlines, all four metrics for the component contrasts, both cutoff continuation rules, and the principal transition and cost findings. The supplement follows those questions with complete tables, protocol details, and qualitative evidence.

| Section | Question and supporting material |
|---|---|
| A | Human evaluation: strict versus tied preferences, task breakdowns, and the study interface. |
| B | Seed robustness: complete per-seed results behind the main-text averages. |
| C | Image quality and detector sensitivity: evaluator and object-level breakdowns. |
| D | Component controls: exchange, calibration, and inactive text; full cells and paired uncertainty. |
| E | Guidance duration: all cutoffs, continuation controls, and matched measurements. |
| F | Transition timing: complete appearance, disappearance, and return curves. |
| G | Computational cost: B200 runtime, peak memory, batching, and host overhead. |
| H | Mechanism diagnostics: denoising-step and boundary behavior. |
| I | Movement controls: exchange–gain interaction and semantic-evaluator validation. |
| J | CogVideoX: authors' suggested CFG 6 and development-selected CFG 3.5. |
| K | Two-object extension: spatial permissions and implementation. |
| L | Mixed-lifecycle benchmark: prompts, schedules, and construction protocol. |
| M | Qualitative evidence: Wan, CogVideoX, and URSA comparisons and residual errors. |

## A HUMAN EVALUATION

**Design and stimuli.** All 30 participants judge the same 18 randomly selected prompts: 6 one-object, 6 two-object, and 6 mixed. Native Wan, TC, and CG supply 54 videos in total. Participants watch all three videos before separately judging timing and quality; equally good videos may be co-selected. Methods are shown under A/B/C labels, with the method-to-label mapping varied across cases.

**Timing assistance.** Following TC (Schiber et al., 2026), helper strips use identical timestamps across methods and mark requested onsets and presence/absence intervals. Participants must still watch the videos. Figures 7–9 show the interface.

### A.1 STRICT AND TIE-INCLUSIVE PREFERENCES

CG is the most frequent participant favorite for both timing (93.3%) and visual quality (50.0%), compared with 3.3% and 26.7% for TC (Table 5). Including ties, 96.7% favor CG for timing and 60.0% for quality. These preferences support ChronoGuide's temporal control and visual quality.

A participant's *strict favorite* is their uniquely most-selected method across cases; *ties included* also counts shared maxima. Rates use all 30 participants. Strict shares plus "No unique favorite" sum to 100%; tie-inclusive shares can exceed 100% because each tied method receives a count.

### A.2 SELECTIONS BY TASK AND CRITERION

Table 6 reports descriptive selections by task, complementing the participant-level favorites above. Each task contains six distinct prompts rated by all thirty participants. Timing and quality are judged separately; co-selections count for every selected method, so rates need not sum to 100%.

Table 5: **Participant preferences.** Count out of 30 (percentage). Strict: sole favorite; ties included: sole or joint favorite. Columns overlap because the latter includes the former. Bold: highest preference rate per column.

| | Timing (out of 30) | | Visual quality (out of 30) | |
|---|---|---|---|---|
| Method | Strict | Ties included | Strict | Ties included |
| Wan2.1-S (native) | 0 (0.0%) | 0 (0.0%) | 2 (6.7%) | 6 (20.0%) |
| Wan2.1-S + TC | 1 (3.3%) | 2 (6.7%) | 8 (26.7%) | 12 (40.0%) |
| Wan2.1-S + CG (ours) | **28 (93.3%)** | **29 (96.7%)** | **15 (50.0%)** | **18 (60.0%)** |
| No unique favorite | 1 (3.3%) | – | 5 (16.7%) | – |

Table 6: Human selections by task and criterion (%; higher means more selections). Each task has six prompts and thirty raters; Overall pools 540 participant–case decisions. Timing and visual quality are separate judgments. Co-selections are allowed, so rates need not sum to 100%. Bold: highest descriptive rate in each column and criterion.

| Criterion | Method | One-object | Two-object | Mixed | Overall |
|---|---|---|---|---|---|
| Timing | Wan2.1-S (native) | 18.9 | 11.7 | 37.2 | 22.6 |
| Timing | Wan2.1-S + TC | 36.1 | 22.2 | 33.3 | 30.6 |
| Timing | Wan2.1-S + CG (ours) | **71.1** | **81.7** | **48.9** | **67.2** |
| Quality | Wan2.1-S (native) | 25.0 | 31.1 | 27.8 | 28.0 |
| Quality | Wan2.1-S + TC | 38.9 | 37.8 | **50.6** | 42.4 |
| Quality | Wan2.1-S + CG (ours) | **61.1** | **51.7** | 34.4 | **49.1** |

Figure 6 summarizes participant favorites across all eighteen prompts. Table 6 gives the task breakdown.

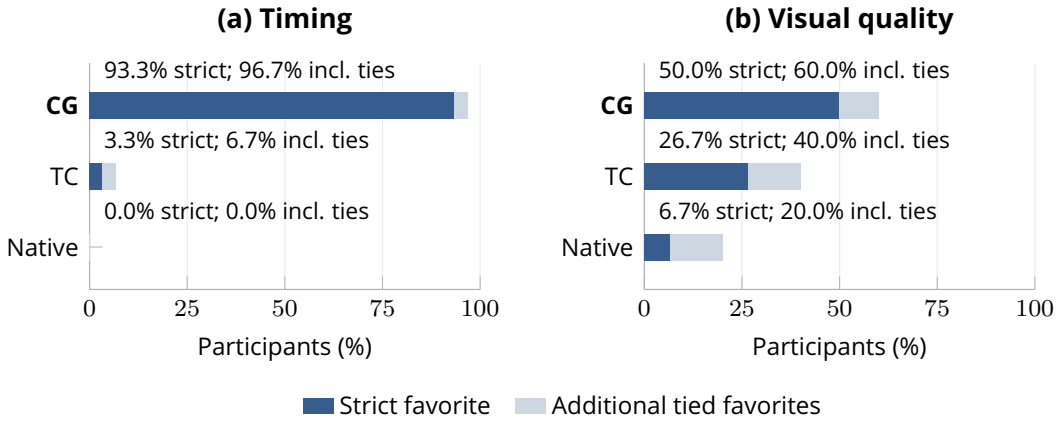



Figure 6: **Participant preferences.** Each participant's favorite is their most-selected method across all cases. Dark bars count strict favorites; pale extensions add shared favorites, giving the tie-inclusive total. All rates use 30 participants; higher bars indicate more support.

### A.3 STUDY INTERFACE AND TIMING AIDS

The study presents viewing instructions, separate timing and quality questions, and timestamped helper strips (Figures 7–9).

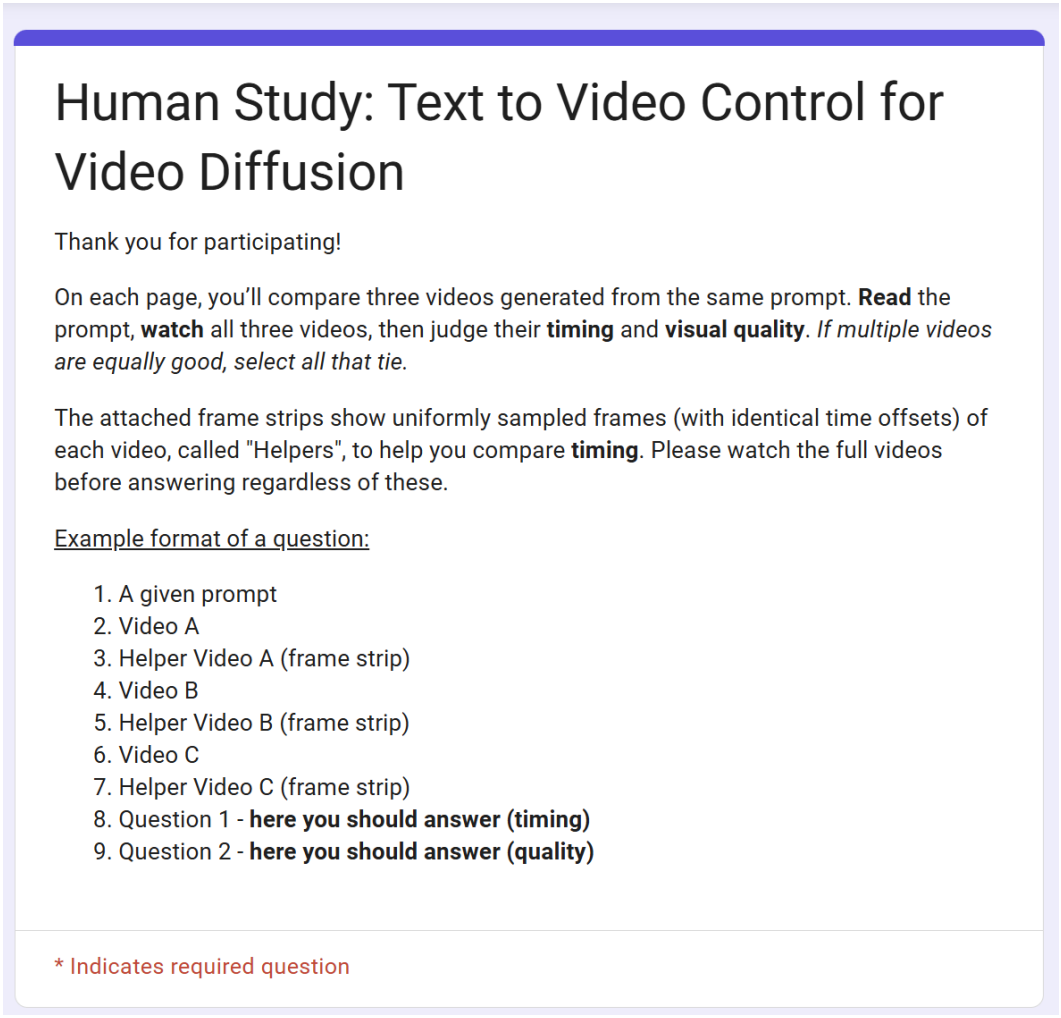



Figure 7: Study instructions, including the requirement to watch the videos, the timestamped helper strips, and the option to select tied preferences.

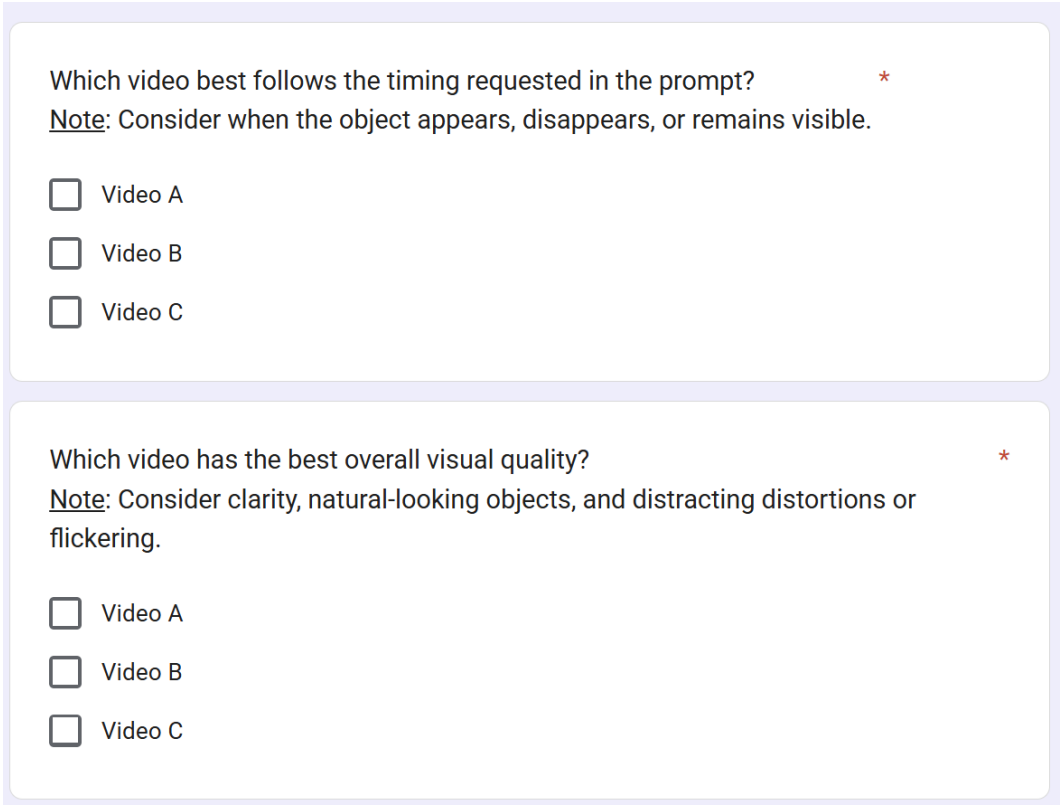



Figure 8: The two study questions. Timing concerns appearance, disappearance, and visibility; visual quality concerns clarity, natural appearance, distortions, and flicker. The checkbox format permits ties.

**Example prompt.** An empty scene. Suddenly, during the second second of the video, a banana appears out of nowhere, drawing all attention. The banana is carried away and is completely out of view by the start of the fourth second of the video.

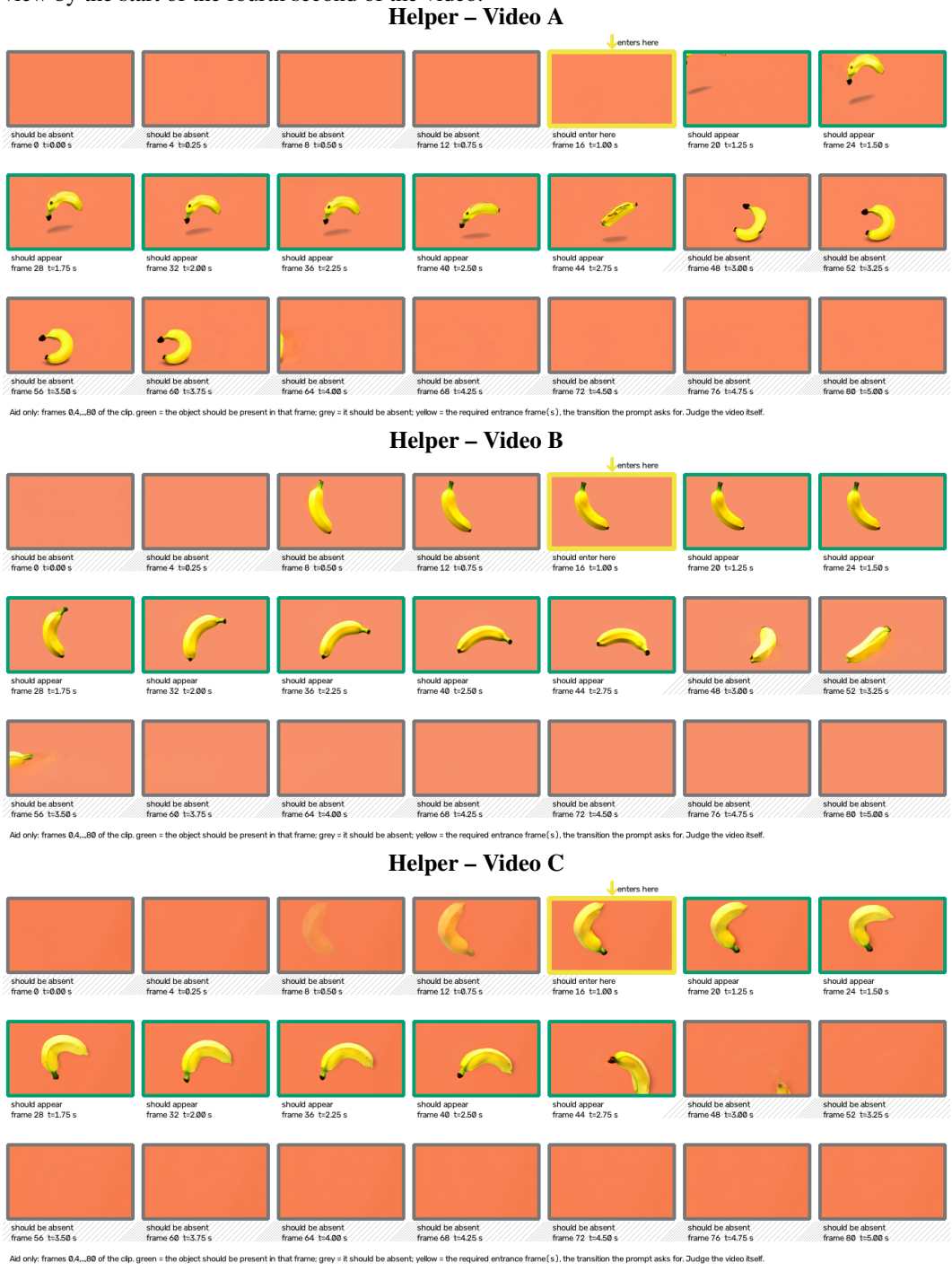



Figure 9: Original helper strips for the mixed banana case shown in the supplied study screenshots: A is native Wan, B is TC, and C is CG. Colors encode the requested schedule, not the method. All methods share timestamps and schedule markers; participants also watch the complete videos. These strips document the study interface rather than a selected perfect result.

## B SEED ROBUSTNESS

**Seed robustness protocol.** Seed 42 matches TC's released benchmark script. We add four random initializations, seeds 43–46, to evaluate five seeds in total with the same configuration. Each covers all 80 prompts with unchanged published-evaluator and MUSIQ scoring.

Table 7: CG (ours) on ONE across seeds 42–46. Mean and sample SD summarize variation across five initializations of the same configuration.

| Seed / summary | TA | Presence | Absence | MUSIQ |
|---|---|---|---|---|
| 42 | 90.06 | 83.88 | 96.25 | 57.12 |
| 43 | 85.81 | 82.50 | 89.13 | 67.21 |
| 44 | 85.88 | 73.38 | 98.38 | 43.47 |
| 45 | 86.31 | 77.25 | 95.38 | 56.47 |
| 46 | 88.50 | 82.13 | 94.88 | 62.04 |
| Mean, 42–46 | 87.31 | 79.83 | 94.80 | 57.26 |
| SD, 42–46 | 1.89 | 4.39 | 3.44 | 8.84 |

The five-seed mean exceeds the published reference on all four metrics. The table reports each seed and the complete five-seed summary.

**Is the timing gain larger than seed variation?** Every tested seed exceeds TC's published TA (83.56) and Abs (87.38). The mean TA gain is 3.75 points, or 1.99 across-seed SDs; the Abs gain is 7.42 points, or 2.15 SDs. The 95% confidence intervals for their means lie above the corresponding published scores (Table 8).

We use one aggregate per seed, giving five independent initializations of the same 80-prompt benchmark. Two-sided one-sample Student $t$ tests compare these means with the fixed published TC values; Holm correction covers all four metrics. Both temporal gains remain significant at 0.05. Inference concerns CG's initialization variability on this benchmark, assuming approximately normal seed means; the published references are held fixed. Presence and MUSIQ have positive mean differences, while their intervals include the reference values. The same-video OWLv2 results independently check detector sensitivity; cross-detector scores are not used for TC significance tests.

Table 8: CG (ours): five-seed inference against fixed published TC scores. CI: 95% Student-$t$ interval for the CG mean; $df = 4$. The last column adjusts all four two-sided tests together.

| Metric | Mean gain | CG mean CI | $t$ | Holm $p$ |
|---|---|---|---|---|
| TA | +3.75 | [84.97, 89.66] | 4.44 | 0.034 |
| Pres | +0.08 | [74.37, 85.28] | 0.04 | 1.000 |
| Abs | +7.42 | [90.52, 99.08] | 4.82 | 0.034 |
| MUSIQ | +0.34 | [46.29, 68.24] | 0.09 | 1.000 |

## C IMAGE QUALITY AND DETECTOR SENSITIVITY

**OWLv2 settings and score aggregation.** We use frozen OWLv2-base-patch16-ensemble (Minderer et al., 2023) with its checkpoint-provided processor. Each object is queried separately as "a {object}", using the lowercased object phrase. A frame counts as present when its highest-scoring postprocessed detection reaches a confidence threshold of 0.15. Videos are decoded sequentially. For P scoring on the 81-frame Wan videos, we sample the indices $\lfloor \mathrm{linspace}(0, 80, 20) \rfloor$ and use the original temporal mask without its first entry, matching the published-evaluator convention. TA averages per-video correctness; Presence and Absence pool correct decisions over requested-present and requested-absent frames, respectively. Seeds 42–46 use the same scorer and checkpoint. Evaluating identical videos isolates detector sensitivity: high TA and Presence under both scorers support control beyond one detector's recognition behavior.

Table 9: CG (ours): OWLv2 on the same complete 80-video panels for each seed. Published-evaluator schedule; frozen OWLv2 query, threshold and checkpoint. MUSIQ repeats Table 7 because the generated videos are identical.

| Seed / summary | TA | Presence | Absence | MUSIQ |
|---|---|---|---|---|
| 42 | 93.25 | 93.13 | 93.38 | 57.12 |
| 43 | 87.31 | 91.00 | 83.63 | 67.21 |
| 44 | 92.19 | 86.25 | 98.13 | 43.47 |
| 45 | 93.31 | 94.50 | 92.13 | 56.47 |
| 46 | 92.94 | 96.38 | 89.50 | 62.04 |
| Mean | 91.80 | 92.25 | 91.35 | 57.26 |
| SD | 2.55 | 3.89 | 5.33 | 8.84 |

### C.1 Full-timeline results across hosts

The two protocols answer different evaluation needs. P reproduces TC's released one-object evaluator: it drops the first control entry, then samples twenty frames uniformly across the clip and pairs them with the twenty remaining entries. Keeping P makes our one-object numbers directly comparable with the published baseline.

F retains every control entry and samples twenty-one aligned frames, including the initial state and both clip endpoints. This is also the convention in TC's released two-object evaluator, whose parser retains the full control sequence. MIX uses F because initial presence, disappearance, and return all form part of its requested lifecycle. For an 81-frame Wan video, F samples indices $0, 4, \ldots, 80$ against the matching 21 control bins; P samples twenty evenly spaced indices against the mask without its first entry. All the methods evaluated by us within each comparison use the same protocol; published baseline scores are retained as reported.

**Published two-object reference.** Our two-object evaluation retains all 21 control entries, following TC's released scorer. TC's published TA equals the arithmetic mean of its reported Pres and Abs, whereas our schedules contain eleven requested-present and ten requested-absent entries. The historical aggregation underlying the published table is not established by the released code; we therefore retain those scores unchanged and identify them as published references.

For TWO, a requested-absent decision is correct only when the persistent object is detected and the entrant is absent; a requested-present decision requires both objects. Thus, removing both objects cannot improve the absence score.

Table 10 reports F on the same CG videos used for the main P comparison. MUSIQ is unchanged. The one-object component check in Table 15 also retains exchange's TA gain under F, showing that the timing benefit persists under both alignments.

Table 10: CG across hosts under full-timeline scoring (F), 80 prompts per host.

| Method | Params / eval. | TA | Pres | Abs | MUSIQ |
|---|---|---|---|---|---|
| CogX-S + CG (ours) | 2B / F | 75.36 | 70.50 | 79.77 | 49.53 |
| Wan2.1-S + CG (ours) | 1.3B / F | 91.13 | 87.00 | 94.89 | 57.12 |
| URSA + CG (ours) | 1.7B / F | 74.11 | 64.88 | 82.50 | 37.90 |

On seed 42, OWLv2 measures 93.25 TA and 93.13 Presence, compared with YOLO's 90.06 and 83.88. Strong control therefore persists with an open-vocabulary detector; neither scorer is treated as human ground truth.

The per-object breakdown separates missed requested appearances (low Presence) from unwanted persistence (low Absence).

Table 11: CG (ours): two detectors on the same 80 main seed 42 videos under published-evaluator. YOLO remains the headline evaluator; OWLv2 is a sensitivity analysis. MUSIQ is detector-independent.

| Detector | TA | Pres | Abs | MUSIQ |
|---|---|---|---|---|
| YOLO | 90.06 | 83.88 | 96.25 | 57.12 |
| OWLv2 | 93.25 | 93.13 | 93.38 | 57.12 |

Table 12: CG (ours), per-object results on all 80 one-object prompts, seed 42, published scoring (P). Each class has one prompt; these scores describe errors rather than establish a class ranking.

| Object | n | TA | Pres | Abs | MUSIQ |
|---|---|---|---|---|---|
| airplane | 1 | 95.00 | 93.75 | 100.00 | 60.20 |
| apple | 1 | 95.00 | 100.00 | 91.67 | 55.11 |
| backpack | 1 | 100.00 | 100.00 | 100.00 | 55.26 |
| banana | 1 | 100.00 | 100.00 | 100.00 | 54.68 |
| baseball bat | 1 | 95.00 | 91.67 | 100.00 | 53.40 |
| baseball glove | 1 | 40.00 | 0.00 | 100.00 | 60.89 |
| bear | 1 | 100.00 | 100.00 | 100.00 | 51.48 |
| bed | 1 | 95.00 | 100.00 | 91.67 | 52.76 |
| bench | 1 | 95.00 | 93.75 | 100.00 | 73.91 |
| bicycle | 1 | 95.00 | 93.75 | 100.00 | 68.93 |
| bird | 1 | 95.00 | 93.75 | 100.00 | 61.41 |
| boat | 1 | 80.00 | 100.00 | 0.00 | 68.93 |
| book | 1 | 100.00 | 100.00 | 100.00 | 52.73 |
| bottle | 1 | 45.00 | 8.33 | 100.00 | 57.96 |
| bowl | 1 | 95.00 | 87.50 | 100.00 | 50.17 |
| broccoli | 1 | 95.00 | 100.00 | 91.67 | 57.42 |
| bus | 1 | 75.00 | 68.75 | 100.00 | 73.59 |
| cake | 1 | 100.00 | 100.00 | 100.00 | 58.06 |
| car | 1 | 75.00 | 93.75 | 0.00 | 62.27 |
| carrot | 1 | 85.00 | 100.00 | 75.00 | 56.82 |
| cat | 1 | 90.00 | 100.00 | 50.00 | 65.01 |
| cell phone | 1 | 100.00 | 100.00 | 100.00 | 46.78 |
| chair | 1 | 90.00 | 75.00 | 100.00 | 53.12 |
| clock | 1 | 100.00 | 100.00 | 100.00 | 49.58 |
| couch | 1 | 95.00 | 100.00 | 91.67 | 59.19 |
| cow | 1 | 95.00 | 93.75 | 100.00 | 65.83 |
| cup | 1 | 95.00 | 87.50 | 100.00 | 50.51 |
| dining table | 1 | 100.00 | 100.00 | 100.00 | 52.01 |
| dog | 1 | 95.00 | 100.00 | 75.00 | 60.80 |
| donut | 1 | 100.00 | 100.00 | 100.00 | 53.19 |
| elephant | 1 | 100.00 | 100.00 | 100.00 | 57.38 |
| fire hydrant | 1 | 95.00 | 93.75 | 100.00 | 67.22 |
| fork | 1 | 100.00 | 100.00 | 100.00 | 52.64 |
| frisbee | 1 | 45.00 | 8.33 | 100.00 | 50.01 |
| giraffe | 1 | 90.00 | 100.00 | 75.00 | 54.33 |
| hair drier | 1 | 80.00 | 0.00 | 100.00 | 47.91 |
| handbag | 1 | 95.00 | 91.67 | 100.00 | 61.48 |
| horse | 1 | 90.00 | 100.00 | 50.00 | 66.18 |
| hot dog | 1 | 60.00 | 0.00 | 100.00 | 54.84 |
| keyboard | 1 | 95.00 | 75.00 | 100.00 | 51.61 |

Table 13: Per-object results (continued), with the same protocol as Table 12.

| Object | n | TA | Pres | Abs | MUSIQ |
|---|---|---|---|---|---|
| kite | 1 | 95.00 | 91.67 | 100.00 | 53.43 |
| knife | 1 | 75.00 | 37.50 | 100.00 | 51.79 |
| laptop | 1 | 100.00 | 100.00 | 100.00 | 50.38 |
| microwave | 1 | 95.00 | 75.00 | 100.00 | 49.46 |
| motorcycle | 1 | 95.00 | 93.75 | 100.00 | 68.30 |
| mouse | 1 | 80.00 | 0.00 | 100.00 | 57.79 |
| orange | 1 | 95.00 | 87.50 | 100.00 | 50.99 |
| oven | 1 | 100.00 | 100.00 | 100.00 | 51.35 |
| parking meter | 1 | 100.00 | 100.00 | 100.00 | 68.06 |
| person | 1 | 90.00 | 100.00 | 50.00 | 70.18 |
| pizza | 1 | 100.00 | 100.00 | 100.00 | 53.58 |
| potted plant | 1 | 95.00 | 87.50 | 100.00 | 58.70 |
| refrigerator | 1 | 95.00 | 75.00 | 100.00 | 52.79 |
| remote | 1 | 100.00 | 100.00 | 100.00 | 48.20 |
| sandwich | 1 | 95.00 | 87.50 | 100.00 | 55.16 |
| scissors | 1 | 100.00 | 100.00 | 100.00 | 47.09 |
| sheep | 1 | 100.00 | 100.00 | 100.00 | 68.66 |
| sink | 1 | 80.00 | 0.00 | 100.00 | 49.33 |
| skateboard | 1 | 95.00 | 91.67 | 100.00 | 58.11 |
| skis | 1 | 40.00 | 0.00 | 100.00 | 55.00 |
| snowboard | 1 | 50.00 | 16.67 | 100.00 | 66.59 |
| spoon | 1 | 90.00 | 100.00 | 83.33 | 56.90 |
| sports ball | 1 | 55.00 | 25.00 | 100.00 | 55.15 |
| stop sign | 1 | 95.00 | 93.75 | 100.00 | 66.06 |
| suitcase | 1 | 100.00 | 100.00 | 100.00 | 62.31 |
| surfboard | 1 | 95.00 | 91.67 | 100.00 | 52.44 |
| teddy bear | 1 | 100.00 | 100.00 | 100.00 | 48.69 |
| tennis racket | 1 | 100.00 | 100.00 | 100.00 | 49.78 |
| tie | 1 | 95.00 | 91.67 | 100.00 | 58.62 |
| toaster | 1 | 80.00 | 0.00 | 100.00 | 54.23 |
| toilet | 1 | 95.00 | 100.00 | 93.75 | 49.45 |
| toothbrush | 1 | 100.00 | 100.00 | 100.00 | 44.06 |
| traffic light | 1 | 95.00 | 93.75 | 100.00 | 65.73 |
| train | 1 | 100.00 | 100.00 | 100.00 | 69.62 |
| truck | 1 | 85.00 | 100.00 | 25.00 | 71.02 |
| tv | 1 | 95.00 | 75.00 | 100.00 | 53.81 |
| umbrella | 1 | 95.00 | 91.67 | 100.00 | 57.11 |
| vase | 1 | 95.00 | 75.00 | 100.00 | 46.50 |
| wine glass | 1 | 95.00 | 87.50 | 100.00 | 51.84 |
| zebra | 1 | 100.00 | 100.00 | 100.00 | 65.50 |

# D COMPONENT ABLATIONS

## D.1 CONTROLLED DESIGN

We evaluate four one-object and six mixed-lifecycle configurations per seed (42/43), each on the same twenty development prompts in two batches of ten: 400 videos in total. Within each task/seed, arms share prompts, schedules, and initialization. ONE uses published scoring (P); MIX uses full-timeline scoring (F). TA averages videos; Pres/Abs pool eligible frame decisions; MUSIQ averages per-video quality.

**Three controlled changes.** *Exchange* compares on versus off while retaining identical temporal text masks. Off is masking only, not native Wan. *Guidance* compares scalar CFG 6 with two-level spatial Haar CFG 6/3/6 (coarse active/inactive/detail), holding exchange and masks fixed. *Inactive text* compares identity-only with identity-plus-event suppression. This tests whether inactive branches must also block event language rather than only object identity.

Table 3 shows all four metrics for the main component contrasts at both seeds. The full factorial cells distinguish the timing contribution of exchange from the quality contribution of calibration.

## D.2 COMPLETE CONFIGURATION RESULTS

Table 14: CG (ours) component study: twenty matched prompts per row, 400 videos total. Compare exchange on/off within a fixed task, seed, mask and CFG; compare Haar/scalar with exchange fixed. Each metric is higher-is-better. Id.: identity; evt: event; Haar: 6/3/6; Scalar: uniform CFG 6. Horizontal rules separate tasks.

| Task | Seed | Mask | Exchange | CFG | TA | Pres | Abs | MUSIQ |
|---|---|---|---|---|---|---|---|---|
| ONE | 42 | Id. | off | Haar | 78.25 | 63.07 | 90.18 | 57.18 |
| ONE | 42 | Id. | on | Haar | 84.75 | 68.18 | 97.77 | 56.55 |
| ONE | 42 | Id. | off | Scalar | 79.75 | 62.50 | 93.30 | 55.86 |
| ONE | 42 | Id. | on | Scalar | 81.75 | 60.23 | 98.66 | 55.38 |
| ONE | 43 | Id. | off | Haar | 78.50 | 62.50 | 91.07 | 62.32 |
| ONE | 43 | Id. | on | Haar | 82.25 | 72.16 | 90.18 | 69.71 |
| ONE | 43 | Id. | off | Scalar | 77.75 | 61.36 | 90.62 | 53.52 |
| ONE | 43 | Id. | on | Scalar | 84.00 | 65.34 | 98.66 | 62.00 |
| MIX | 42 | Id. | off | Haar | 57.14 | 67.60 | 49.38 | 61.11 |
| MIX | 42 | Id. | on | Haar | 85.71 | 81.01 | 89.21 | 58.71 |
| MIX | 42 | Id.+evt | off | Haar | 55.95 | 78.77 | 39.00 | 62.82 |
| MIX | 42 | Id.+evt | on | Haar | 83.10 | 82.68 | 83.40 | 61.20 |
| MIX | 42 | Id.+evt | off | Scalar | 69.29 | 74.86 | 65.15 | 59.65 |
| MIX | 42 | Id.+evt | on | Scalar | 83.57 | 67.04 | 95.85 | 53.84 |
| MIX | 43 | Id. | off | Haar | 59.52 | 46.37 | 69.29 | 63.14 |
| MIX | 43 | Id. | on | Haar | 75.00 | 51.40 | 92.53 | 62.13 |
| MIX | 43 | Id.+evt | off | Haar | 52.38 | 43.02 | 59.34 | 64.13 |
| MIX | 43 | Id.+evt | on | Haar | 74.05 | 53.63 | 89.21 | 66.92 |
| MIX | 43 | Id.+evt | off | Scalar | 55.24 | 44.69 | 63.07 | 60.95 |
| MIX | 43 | Id.+evt | on | Scalar | 75.00 | 58.66 | 87.14 | 61.48 |

## D.3 PAIRED EFFECTS AND UNCERTAINTY

Figures 10–11 show all contrasts and 95% paired prompt-bootstrap intervals. We resample twenty matched prompt pairs 2,000 times within each seed and recompute the aggregate difference, including pooled Pres/Abs. Intervals quantify prompt variation within a seed, not uncertainty across seeds. A zero-crossing interval leaves the effect's direction uncertain.

*Exchange P/F* is on minus off, with Haar fixed. I/E denotes identity-only/event-aware masking on MIX. *Interaction* subtracts the identity-only exchange effect from the event-aware effect. *Haar*

*off/on* is Haar minus scalar CFG with exchange held off/on; MIX uses event-aware masks. Positive effects favor the first configuration in each subtraction.

With Haar fixed, exchange improves TA in all six task/mask/seed comparisons, including 15.48–28.57-point gains on MIX. Its role is to improve schedule adherence; its isolated MUSIQ effect is not uniformly positive. Haar increases MUSIQ in all four exchange-enabled task/seed pairs. Adding event masking with exchange raises MIX Pres/MUSIQ at both seeds, with a 0.95–2.62-point TA and 3.32–5.81-point Abs cost. These contrasts identify complementary roles for exchange, calibration and event masking in the final configuration.

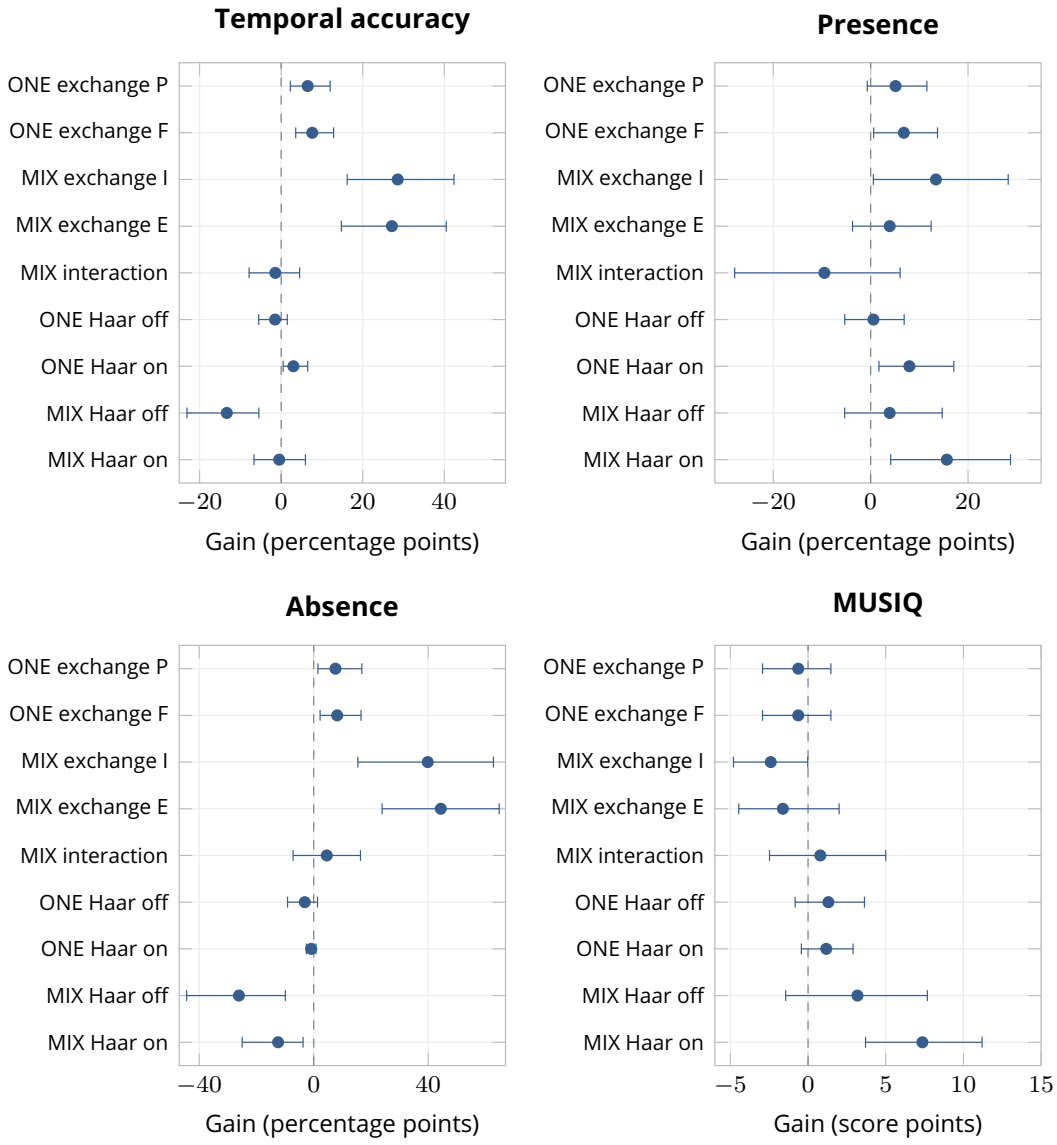



Figure 10: Component attribution, seed 42: exchange controls timing; Haar calibrates quality. Dots are aggregate differences; bars are 95% paired prompt-bootstrap intervals. The dashed zero line marks no change. All four metrics and every contrast are shown; contrast definitions are given above. Temporal units are percentage points; MUSIQ uses score points.

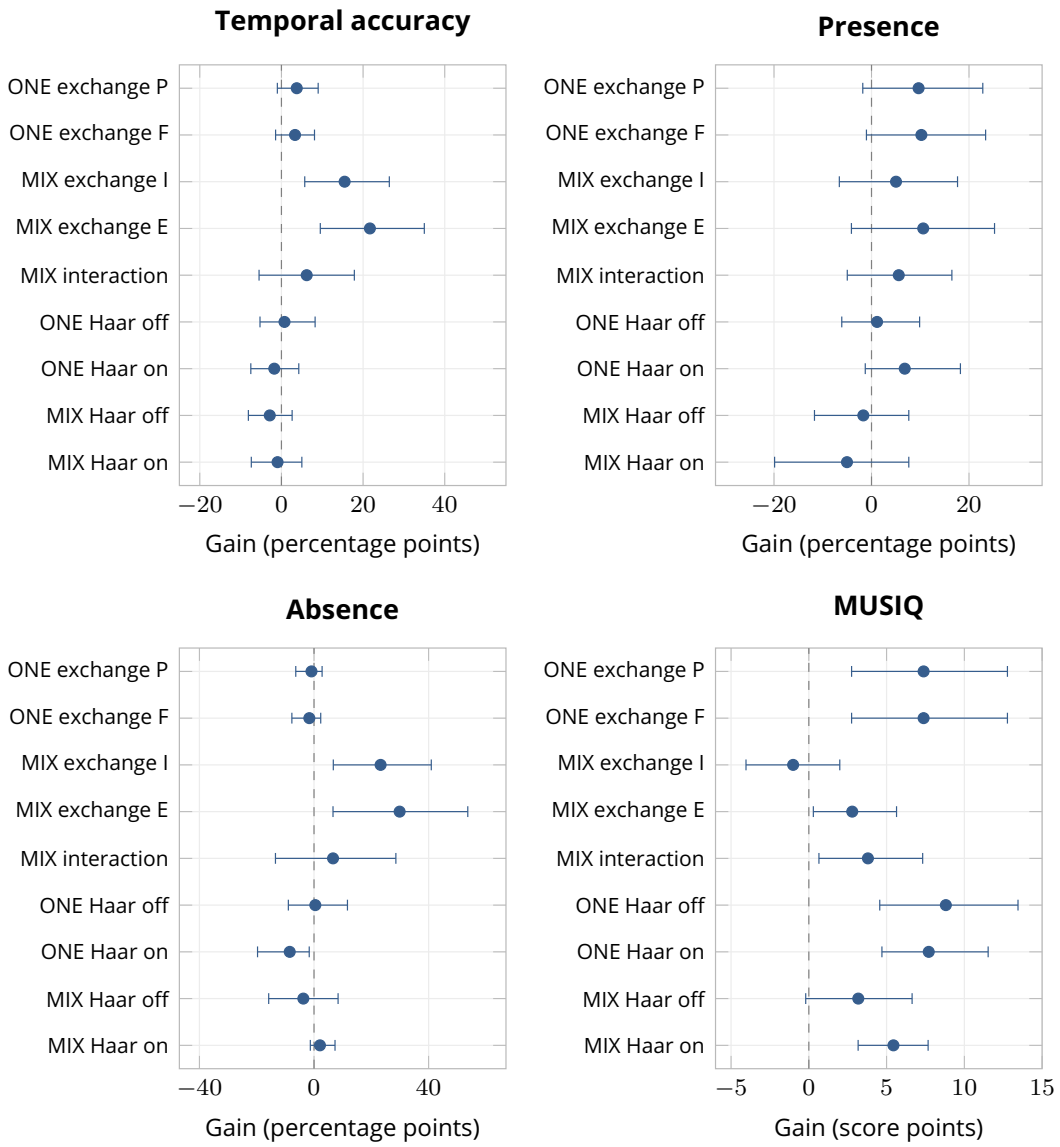



Figure 11: Replication at seed 43: the same component contrasts under another initialization. Dots are aggregate differences; bars are 95% paired prompt-bootstrap intervals. The dashed zero line marks no change. All four metrics and every contrast are shown; contrast definitions are given above. Temporal units are percentage points; MUSIQ uses score points.

### D.4 SCORING-PROTOCOL CHECK

The one-object exchange control can also be scored over all 21 bins (F). The same videos and MUSIQ values are used; this is an evaluator check, not an additional experiment. Both seeds retain positive aggregate TA gains.

Table 15: CG (ours): exchange beyond temporal prompt masking, evaluated on twenty matched one-object development prompts per arm and seed (F). Both arms use Haar CFG 6/3/6; off retains the masks but removes value exchange. MUSIQ uses the same videos as the P results above.

| Seed | Exchange | TA | Pres | Abs | MUSIQ |
|---|---|---|---|---|---|
| 42 | Off | 78.33 | 63.64 | 88.93 | 57.18 |
| 42 | On | 85.95 | 70.45 | 97.13 | 56.55 |
| 43 | Off | 78.57 | 62.50 | 90.16 | 62.32 |
| 43 | On | 81.90 | 72.73 | 88.52 | 69.71 |

### D.5 MECHANISTIC ROUTING CONTROLS

**Controlled design.** The component ablation tests whether value exchange helps beyond temporal text masking. Here, three matched controls test why the routing rule matters. *Wrong-state donors* use the incompatible state-conditioned read at the same exchange locations. *Cross-state blocking* removes connections between different requested states and renormalizes the remaining weights. *Global replacement* applies receiver-conditioned values to all connections, including same-state pairs. CG instead retains live values within a state and matches cross-state values to the receiver.

The first three controls generate 240 videos on twenty development prompts per task (ONE/MIX), seeds 42/43, and three controls. All twelve planned cells are complete. The controls share 80 reused CG (ours) baseline videos, with prompts, schedules, initialization, noise, model settings, batch composition and evaluator matched. Fixed batches contain ten videos; batch size is separate from guidance duration. ONE uses P scoring and MIX uses F. Baseline reuse does not add independent observations, and no results are selected by outcome.

**State, connectivity and selectivity.** CG has higher mean TA in all twelve comparisons. Wrong-state donors test receiver-state compatibility: their TA deficit is 5.25–24.76 points. Blocking reduces TA by 9.05–17.50 points and required presence by 35.75–53.07. Its higher absence in three cells accompanies substantially lower required presence, supporting retention of cross-state communication. Global replacement reduces mean TA by 1.00–16.75 points, supporting preservation of live values on same-state connections for timing.

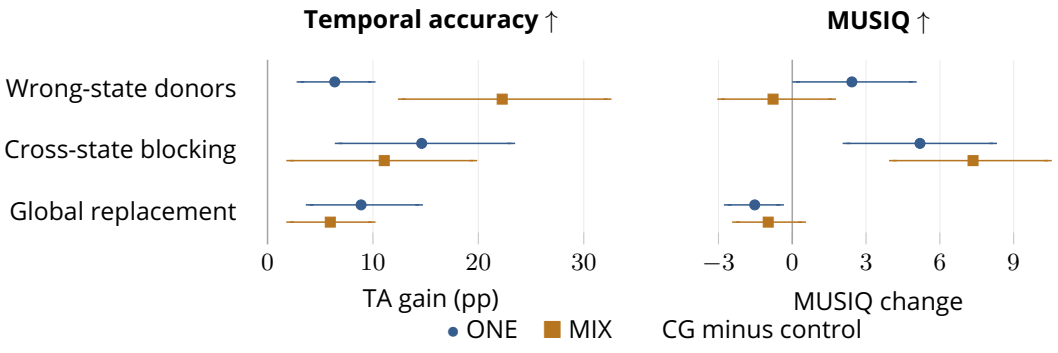



Figure 12: **Receiver-matched routing improves timing.** CG minus control on twenty prompts/task, seeds 42/43. Bars: 95% paired prompt-bootstrap intervals (2,000 resamples; seeds kept together). Positive favors CG. Appendix D.5.

**Aggregation and uncertainty.** Figure 12 averages each paired difference equally over twenty prompts and the two fixed seeds within a task. Its 2,000 bootstrap resamples sample prompts, retaining both seed observations together. This is a descriptive average over seeds 42/43, not uncertainty over arbitrary new seeds. Figure 13 and Tables 16–17 retain every separate-seed comparison. Their original 95% intervals resample twenty prompt pairs within a task/seed cell. Pres/Abs recompute pooled eligible-frame accuracy in each resample; TA and MUSIQ average videos. Intervals are not adjusted for multiple comparisons. All differences below are CG minus control, so positive values favor CG.

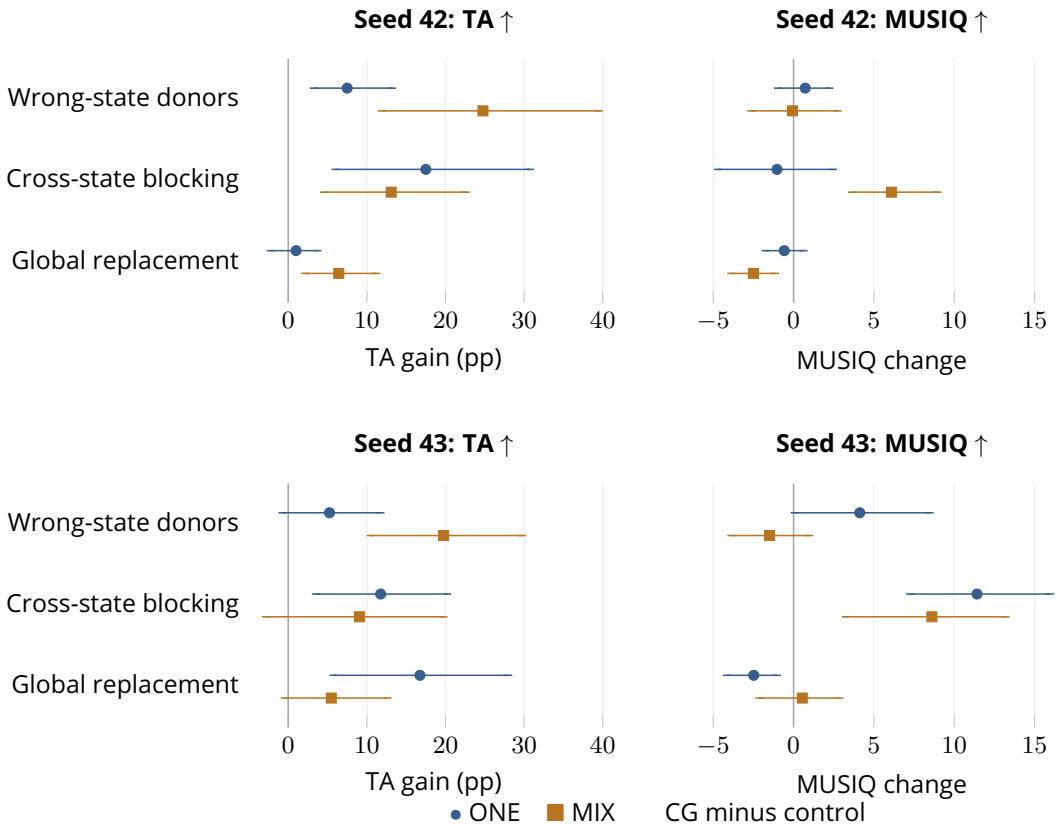



Figure 13: **Routing effects across tasks and seeds.** CG minus control; positive values favor CG. Bars show the original 95% paired prompt-bootstrap intervals within each complete twenty-prompt cell. All twelve comparisons are shown. Colors and markers match Figure 12. Wrong-state donors and renormalized blocking test state compatibility and connectivity; global replacement tests extending substitution to same-state pairs.

**Timing and image quality.** Eight of twelve within-cell TA intervals exclude zero: three of four for wrong-state donors, three for blocking, and two for global replacement. All four blocking Presence intervals exclude zero. Global replacement improves MUSIQ in MIX 42 and ONE 43 with intervals excluding zero, while lowering TA; the other two MUSIQ differences are uncertain. Global replacement changes same-state communication as well as cross-state reads. Its two-seed mean MUSIQ gains of 1.53/0.98 on ONE/MIX accompany TA losses of 8.88/5.95. We retain selective replacement because the objective is adherence to the requested temporal schedule; frame quality alone cannot establish correct timing. These interventions isolate the routing rule's contribution to temporal adherence. The complete method improves both TA and MUSIQ over same-host TC and native generation in the headline comparisons (Tables 1 and 2). MUSIQ measures image quality rather than temporal coherence or action identity.

Table 16: Routing controls on ONE: CG (ours) minus control, twenty matched prompts per seed. Positive differences favor CG. Each cell gives the difference and its within-seed 95% paired prompt-bootstrap interval. Temporal metrics use percentage points; MUSIQ uses score points.

| Control | ΔTA | ΔPres | ΔAbs | ΔMUSIQ |
|---|---|---|---|---|
| Seed 42 | | | | |
| Wrong-state donors | +7.50 [3.25, 13.25] | +9.09 [1.67, 17.22] | +6.25 [0.00, 16.52] | +0.73 [−0.98, 2.24] |
| Cross-state blocking | +17.50 [6.00, 30.75] | +36.93 [13.51, 59.31] | +2.23 [−1.92, 9.90] | −1.03 [−4.71, 2.46] |
| Global replacement | +1.00 [−2.25, 3.75] | +0.00 [−7.05, 6.40] | +1.79 [0.00, 5.10] | −0.57 [−1.74, 0.64] |
| Seed 43 | | | | |
| Wrong-state donors | +5.25 [−0.75, 11.75] | +11.93 [−1.67, 26.14] | +0.00 [−6.37, 4.74] | +4.12 [0.05, 8.48] |
| Cross-state blocking | +11.75 [3.50, 20.25] | +39.20 [22.55, 56.98] | −9.82 [−21.11, −2.97] | +11.42 [7.24, 15.96] |
| Global replacement | +16.75 [5.75, 28.00] | −1.14 [−7.98, 5.11] | +30.80 [12.28, 51.36] | −2.48 [−4.15, −1.03] |

Table 17: Routing controls on MIX: CG (ours) minus control, twenty matched prompts per seed. Positive differences favor CG. Each cell gives the difference and its within-seed 95% paired prompt-bootstrap interval. Temporal metrics use percentage points; MUSIQ uses score points.

| Control | ΔTA | ΔPres | ΔAbs | ΔMUSIQ |
|---|---|---|---|---|
| Seed 42 | | | | |
| Wrong-state donors | +24.76 [11.90, 39.52] | +13.41 [0.00, 30.92] | +33.20 [10.31, 56.02] | −0.07 [−2.65, 2.75] |
| Cross-state blocking | +13.10 [4.52, 22.62] | +53.07 [36.04, 69.68] | −16.60 [−30.49, −6.86] | +6.09 [3.62, 8.96] |
| Global replacement | +6.43 [2.14, 11.19] | +5.03 [−3.91, 16.45] | +7.47 [−1.08, 17.04] | −2.50 [−3.89, −1.15] |
| Seed 43 | | | | |
| Wrong-state donors | +19.76 [10.48, 29.76] | +11.73 [−2.80, 25.29] | +25.73 [7.01, 44.80] | −1.49 [−3.88, 0.97] |
| Cross-state blocking | +9.05 [−2.86, 19.76] | +35.75 [17.06, 52.66] | −10.79 [−26.40, −0.40] | +8.61 [3.25, 13.21] |
| Global replacement | +5.48 [−0.48, 12.62] | +5.59 [−12.94, 24.12] | +5.39 [0.00, 13.90] | +0.54 [−2.14, 2.86] |

### D.6 TWO-OBJECT EXCHANGE CONTROL

We add forty control videos on the complete two-object development panel, paired with forty archived CG baselines at seed 42. This brings the four mechanism controls to 280 new videos and 120 reused baselines. The two-object panel uses batches of eight and fifty solver updates. Prompts, schedules, accepted initialization, model settings and the original evaluator are matched within the study. These development scores are separate from the held-out benchmark in Table 1.

The control removes only cross-state value corrections. Regional text permissions, mixed retrieval, same-state positive replacement, source/target text gain, and donor streams remain active. It therefore measures the incremental contribution of cross-state exchange beyond the composition controls in Appendix K, rather than disabling the entire two-object extension.

Table 18: Two-object exchange control: forty paired development prompts, seed 42. All metrics ↑. Differences are CG minus exchange off; temporal differences use percentage points. Intervals use 2,000 paired prompt-bootstrap resamples (95% percentile, not multiplicity-adjusted).

| Metric | CG (ours) | Exchange off | Difference | 95% interval |
|---|---|---|---|---|
| Joint-state TA | 61.79 | 55.36 | +6.43 | $[-2.50, 16.55]$ |
| Required presence | 52.27 | 50.91 | +1.36 | $[-10.45, 14.55]$ |
| Required absence | 72.25 | 60.25 | +12.00 | $[-0.50, 24.75]$ |
| Incumbent presence | 81.19 | 80.71 | +0.48 | $[-7.98, 9.88]$ |
| MUSIQ | 71.50 | 70.94 | +0.56 | $[-1.01, 2.23]$ |

Full CG has higher means on all five metrics, with the largest changes in required absence and joint-state accuracy. The incumbent's mean presence is nearly unchanged, while mean MUSIQ is slightly higher. These results are consistent with an additional temporal-control benefit beyond regional permissions and text amplification. All five intervals include zero, so the panel provides directional evidence rather than a conclusive effect or a formal quality-equivalence test.

## E GUIDANCE DURATION

### E.1 DURATION AND CONTINUATION CONTROLS

C$k$ applies CG for the first $k$ of fifty updates. Every arm executes all fifty solver updates with the same timesteps, solver history, and random stream. After cutoff, *native* continuation uses ordinary Wan CFG 6 and the original prompt. *Conditioning-only* continuation keeps temporal text masks and Haar guidance but stops value exchange. C50 is full CG. Thus these controls distinguish the duration of exchange from the duration of temporal conditioning itself.

Figures 4 and 14 report the four metrics under both continuation rules on twenty development prompts per task, seed 42. C10 improves ONE TA/MUSIQ by 1.75/0.51 points with unchanged Abs. It therefore offers a useful fast setting for simple appearance schedules: apply CG early, then finish with native conditioning. On these development panels, this preserves the timing benefit while reducing intervention cost. For MIX, full guidance instead gains 3.57 TA and 3.73 Abs over C10; C10 gains 0.65 MUSIQ. Exits and returns benefit from continued state-specific exchange, and retaining masks/Haar alone does not consistently close the gap. We retain full guidance for the headline comparisons and mixed lifecycles, while C10 is a deployment option for single-object appearance when latency matters.

### E.2 MATCHED ENDPOINT AND SAMPLING COST

For a separate endpoint check, we use the same first ten prompts for all cutoffs, including native C0. This matched ten-prompt subset complements the twenty-prompt duration curves above. Synchronized sampling time excludes setup, validation, decoding, and scoring, and is divided by batch size ten. Each timed setting has one batch observation. C10 saves 52% relative to C50; C25 saves about 32%, and C40 about 13%.

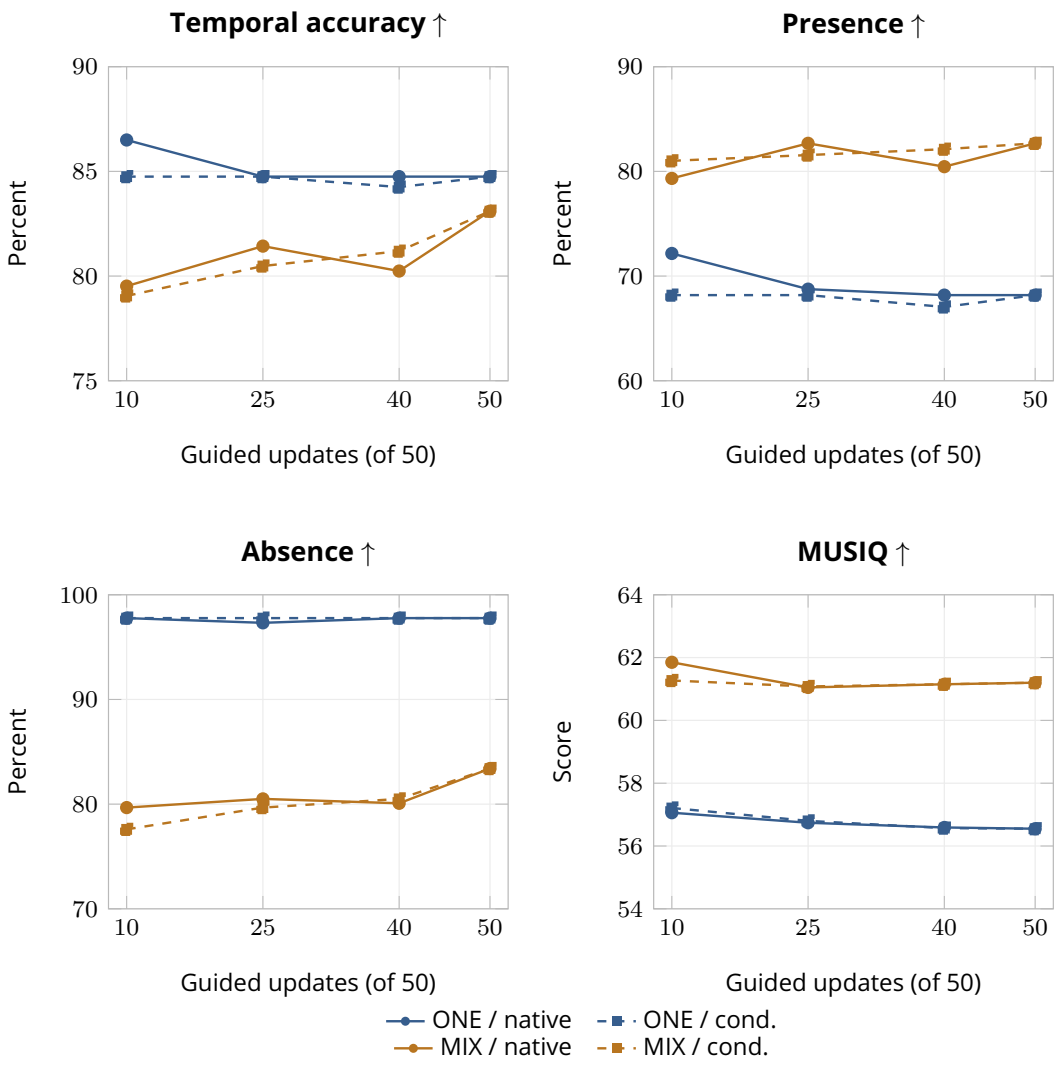



Figure 14: Full duration ablation: twenty development prompts per task, seed 42. Blue: ONE (P); gold: MIX (F). Solid/circles: native continuation; dashed/squares: conditioning-only continuation (cond.). C50 is the shared full-guidance endpoint. All trajectories complete fifty solver updates; only the intervention duration changes. ↑: higher is better.

Table 19: Cutoff versus full guidance on the same ten development prompts per task, seed 42, batch size ten, B200. ONE uses P and MIX uses F. C0 uses native Wan; C50 uses full CG (ours); C10/C25/C40 use CG (ours) until the cutoff, then ordinary Wan CFG. Seconds/video divide synchronized batch sampling time by ten; savings use C50. C0 has no comparable sampling timer (—). Each timed cell has one batch observation.

| Panel | Guidance | TA | Pres | Abs | MUSIQ | s/video | Saving |
|---|---|---|---|---|---|---|---|
| ONE | Native (C0) | 78.00 | 64.77 | 88.39 | 56.88 | — | — |
| ONE | Full (C50) | 93.00 | 85.23 | 99.11 | 54.96 | 194.19 | — |
| ONE | C10 | 96.00 | 92.05 | 99.11 | 55.79 | 92.41 | 52.41% |
| ONE | C25 | 93.00 | 85.23 | 99.11 | 55.15 | 131.12 | 32.48% |
| ONE | C40 | 93.00 | 85.23 | 99.11 | 54.99 | 168.93 | 13.01% |
| MIX | Native (C0) | 44.76 | 64.47 | 33.58 | 59.39 | — | — |
| MIX | Full (C50) | 84.29 | 85.53 | 83.58 | 63.53 | 194.22 | — |
| MIX | C10 | 78.57 | 80.26 | 77.61 | 64.25 | 92.68 | 52.28% |
| MIX | C25 | 80.48 | 85.53 | 77.61 | 63.21 | 130.61 | 32.75% |
| MIX | C40 | 80.00 | 84.21 | 77.61 | 63.47 | 168.74 | 13.12% |

**Complete twenty-prompt timings.** Mean seconds per batch of ten for C10/C25/C40/C50 are 924.2/1309.4/1687.9/1942.0 (ONE, native continuation) and 926.0/1307.5/1686.9/1942.2 (MIX, native). Conditioning-only C10/C25/C40 costs 960.2/1330.7/1696.4 (ONE) and 961.6/1331.2/1696.3 (MIX). These average two batches of different prompts, not repeated timing trials. Appendix G separates these amortized sampling costs from batch-one pipeline latency.

# F Timing of object appearances and disappearances

## F.1 Mixed-lifecycle transitions

The main text reports the disappearance comparison, where CG has its largest advantage. Figure 15 includes every transition type: 27 first appearances, 26 returns, and 80 exits from the same 80 mixed videos, seed 42. At each relative time, correctness uses the requested state in that bin: detection for presence, no detection for absence. Consequently, a post-return detection is not proof that the preceding disappearance succeeded. All comparisons use the same prompts and F scoring; TC uses $\ell = 5$.

## F.2 One-object onset response

Figure 16 instead plots detection probability around one-object onset. The requested step makes early detections and missed appearances explicit. Here the vertical axis measures detection probability; the mixed-lifecycle plot above measures agreement with the requested state.

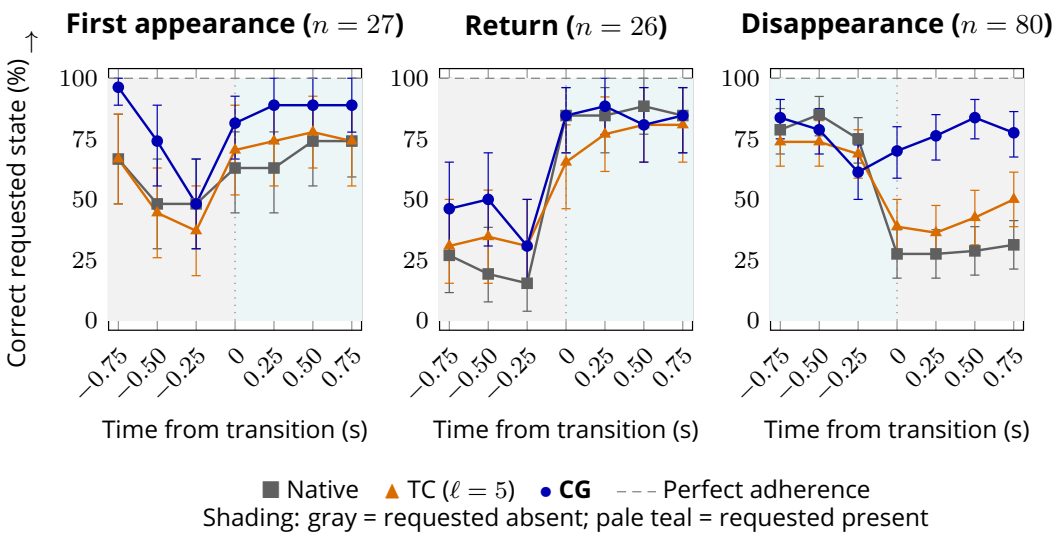



Figure 15: How often is the requested visibility state correct around a lifecycle change? Higher is better throughout: a correct frame has a detection when presence is requested (teal background), and no detection when absence is requested (gray). Zero marks the requested transition; the surrounding points are sampled bins before and after it; each bin spans 0.25 seconds. The dashed line is perfect adherence. Curves cover all 27 first appearances, 26 returns and 80 disappearances in the matched mixed benchmark, seed 42 (F); bars show 95% prompt-bootstrap intervals. ChronoGuide's clearest gain is disappearance: three bins after exit, absence is correct in 77.5% of events, versus 50.0% for TC and 31.25% for native Wan. Return remains harder because the object must first disappear and then reappear. Native Wan uses ordinary CFG; TC uses $\ell = 5$.

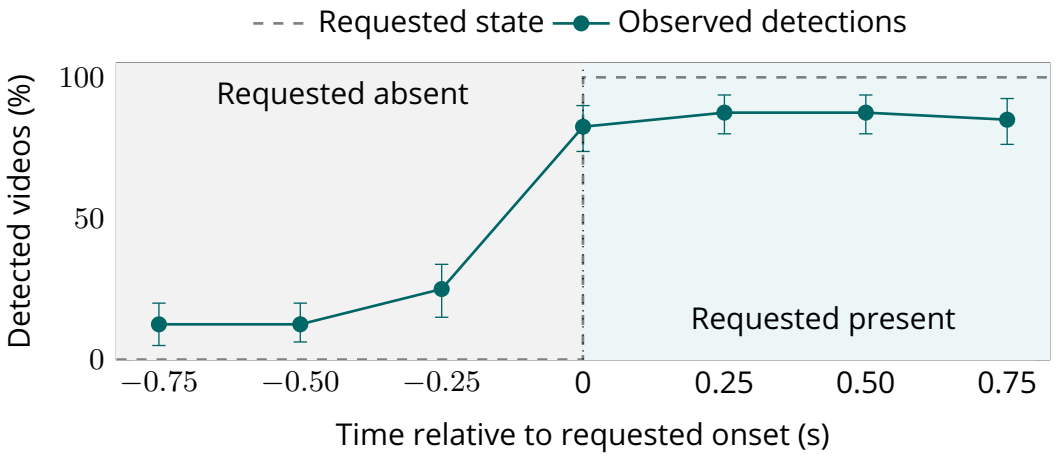



Figure 16: One-object boundary response on the same 80 videos (F). Zero is the first sampled bin in which presence is requested; each bin spans 0.25 seconds. The solid curve is the fraction of videos with a detection; the dashed step is the requested state. Bars are prompt-bootstrap 95% intervals. Early detections and missed appearances remain visible. Better adherence follows the requested step: lower before onset, higher after.

# G RUNTIME AND MEMORY

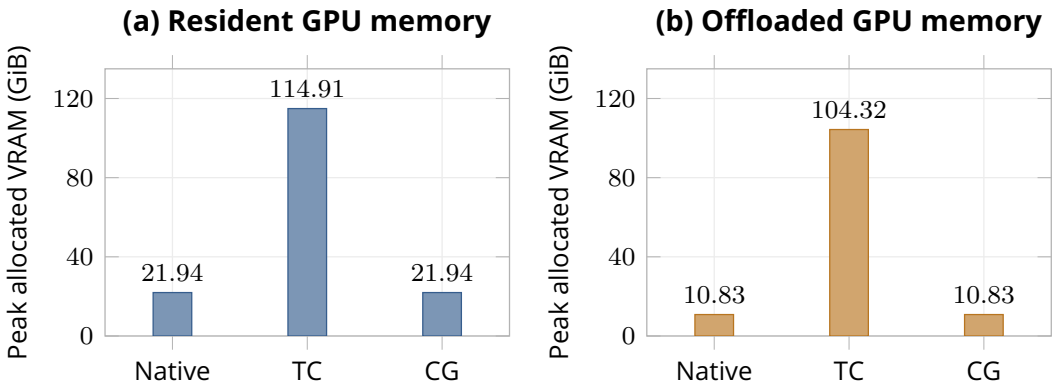



Figure 17: **Peak GPU memory on B200.** Wan 2.1-S, batch one, TC $\ell = 5$. Lower memory use is better. Resident/offloaded weights are compared separately. CG matches native Wan's peak allocation and uses 81%/90% less than local TC. TC exceeds 100 GiB in both measured configurations, requiring substantially more GPU capacity. These are local measurements; complete latency comparisons appear below.

## G.1 COMPUTATIONAL COMPLEXITY

For $L$ video tokens and channel width $d$, dense attention has $O(L^2 d)$ computation per block. A fixed number of donor streams and state-conditioned value reads changes the constant factor, not this order. The donor value tensors add $O(Ld)$ storage per block; chunked reads avoid materializing a separate dense matrix for every receiver-state pair. Auxiliary streams do not maintain independent sampling trajectories. The same operation precedes either continuous predictions or categorical logits; the host's solver or discrete refinement rule is unchanged. Actual peak memory also includes weights, activations, attention workspaces, and the allocator, which we measure separately below.

## G.2 SINGLE-VIDEO LATENCY AND PUBLISHED REFERENCE

We measure CG with resident weights on a single NVIDIA B200: two fixed one-object development prompts, seed 42, batch one, one warm-up and three measured calls per prompt/configuration. Every configuration generates 81 frames at $480 \times 832$ using fifty solver updates. The synchronized timer includes text encoding, sampling, and VAE decoding; checkpoint loading and preflight validation are excluded. The same physical GPU is used across configurations within each prompt.

Table 20: Single-video latency on B200 (seconds; lower is better). CG: mean $\pm$ repetition SD and mean across two prompts, measured with a common pipeline timer. TC: author-reported latency (Schiber et al., 2026), separated by source. Percentage reductions compare the measured CG configurations.

| Method | Carrot | Dog | Mean |
|---|---|---|---|
| CG C50 (ours) | $213.391 \pm 0.083$ | $213.282 \pm 0.088$ | 213.34 |
| CG C25 (ours) | $141.683 \pm 0.017$ | $141.588 \pm 0.097$ | 141.64 |
| CG C10 (ours) | $98.662 \pm 0.021$ | $98.604 \pm 0.030$ | 98.63 |
| TC (author-reported) | Single B200; five guided steps | | $\sim 307$ |

**Timing protocol and comparison scope.** For the published benchmark, we take TC's accuracy and latency from the authors' report, preserving its reported operating point. Table 20 identifies the timing source for each method. Our CG configurations use the same physical B200 within

each prompt, resident weights, archived seed-42 initialization, output dimensions, solver updates and timer boundaries. This comparison isolates the cost of guidance duration. A shared GPU model alone does not control software and timer differences across papers, so latency reductions are computed within this measured series.

C25 and C10 reduce latency by 33.6% and 53.8% relative to C50. Guidance ends after 25 or 10 updates, followed by native continuation through all fifty solver updates. The separate twenty-prompt duration study (Appendix E) supplies the quality and temporal-accuracy evidence: ONE benefits from short guidance, while mixed lifecycles favor continued exchange. Thus intervention duration provides a measured efficiency choice without reducing the generation step budget.

Removing instrumentation overhead while preserving the accepted final latent tensors gives a further C50 implementation measurement of 207.16 s. Table 20 consistently reports the original configuration. Earlier loaded-weight pipeline measurements were 69.94 s for native Wan and 214.87 s for full CG; these are separate from the new resident cutoff study. Batched sampling costs below have a different scope and are not single-video latency.

### G.3 GPU VRAM ACROSS BATCH AND RESIDENCY SETTINGS

Each GPU-memory measurement runs one method in isolation on B200 after warm-up. Resident keeps the model weights on the GPU; Offloaded moves model weights to host memory when they are not in use. Batch-one entries take the maximum over two prompts per task. The mixed batch-ten comparison uses one call on the same ten prompts for native Wan and ChronoGuide. TC uses $\ell = 5$ and is measured at batch one only.

Table 21: Peak GPU VRAM (GiB), with matching residency within each comparison. Allocated denotes live tensor storage; reserved denotes memory held by the GPU allocator. These GPU counters do not measure CPU/host RAM. These local measurements are separate from published latency numbers.

| Panel / batch | Residency | Method | Allocated | Reserved |
|---|---|---|---|---|
| ONE / 1 | Resident | Wan2.1-S (native) | 21.94 | 24.30 |
| ONE / 1 | Resident | Wan2.1-S + TC | 114.91 | 116.29 |
| ONE / 1 | Resident | Wan2.1-S + CG (ours) | 21.94 | 24.01 |
| ONE / 1 | Offloaded | Wan2.1-S (native) | 10.83 | 11.58 |
| ONE / 1 | Offloaded | Wan2.1-S + TC | 104.32 | 105.74 |
| ONE / 1 | Offloaded | Wan2.1-S + CG (ours) | 10.83 | 11.50 |
| MIX / 1 | Resident | Wan2.1-S (native) | 21.94 | 24.30 |
| MIX / 1 | Resident | Wan2.1-S + TC | 114.91 | 116.35 |
| MIX / 1 | Offloaded | Wan2.1-S (native) | 10.83 | 11.58 |
| MIX / 1 | Offloaded | Wan2.1-S + TC | 104.32 | 105.74 |
| MIX / 10 | Resident | Wan2.1-S (native) | 33.74 | 37.77 |
| MIX / 10 | Resident | Wan2.1-S + CG (ours) | 44.61 | 52.11 |
| MIX / 10 | Offloaded | Wan2.1-S (native) | 22.69 | 27.78 |
| MIX / 10 | Offloaded | Wan2.1-S + CG (ours) | 33.55 | 41.84 |

For one-object generation at batch one, adding ChronoGuide leaves native Wan's measured peak allocated VRAM unchanged: 21.94 GiB with resident weights and 10.83 GiB with offloading. These peaks are 81% and 90% lower than TC under the corresponding residency settings. The offloaded footprint is below the 24 GB capacity of many commercial GPUs. These B200 measurements establish the memory advantage. The mixed batch-ten call uses additional activation storage: allocated memory rises from 33.74 to 44.61 GiB resident and from 22.69 to 33.55 GiB offloaded.

**Memory headroom and throughput.** The smaller batch-one allocation leaves room for batched execution. ChronoGuide supports batches of ten on B200, using 44.61 GiB resident or 33.55 GiB offloaded, demonstrating headroom for batched execution. For a batch of $B$ videos completed in $T_B$ seconds, $T_B/B$ is the amortized cost per video and $T_B$ is the response time for the batch. We use the former below and reserve batching-speedup claims for matched batch-one measurements.

### G.4 CUTOFF COST WITHIN THE SAMPLING PHASE

Table 19 pairs the cutoff quality results with their sampling costs. Every trajectory still executes all fifty solver updates. Timers sum the synchronized sampling updates and exclude setup, validation, decoding and export. Full guidance takes 194.19–194.22 s per video in these batches. C10 lowers this to 92.41–92.68 s, saving 52.28–52.41%; C25 saves 32.48–32.75%, and C40 saves 13.01–13.12%. ONE C10 improves TA and MUSIQ over full guidance; MIX retains its highest TA with full guidance. The four metrics and time together make the deployment choice explicit: shorter intervention can favor throughput and quality, while longer intervention can better enforce mixed lifecycles.

All ten videos finish together, so the reported seconds per video measure amortized sampling cost. TC's published latency has a different timing scope and is reported separately. The local arms share the recorded GPU type and software; each setting has one timing observation.

### G.5 BATCHING AND AMORTIZED COST

Table 22: B200, batch size ten, with the same mixed-lifecycle prompts for both methods. One timed batch follows warm-up. Seconds per video divide batch time by ten; all ten outputs finish together. Variability cannot be estimated from one batch.

| Method | Repeats | Batch time (s) | Seconds / video | SD (s) |
|---|---|---|---|---|
| Wan2.1-S (native) | 1 | 686.76 | 68.68 | – |
| Wan2.1-S + CG (ours) | 1 | 1956.66 | 195.67 | – |

The equal-batch comparison measures throughput for native Wan and full ChronoGuide. It is separate from the batch-one TC reference.

### G.6 TRANSFER-HOST OVERHEAD

Table 23: Sampling plus decoding on B200, batch size one, for the same one-object development prompt on each host. Each entry is one measurement with full guidance; model loading, setup and file export are excluded. Ratios use the native host in the same timing scope. Repeated measurements are needed to estimate variability.

| Method | Seconds | Relative to native |
|---|---|---|
| URSA (native) | 37.46 | — |
| URSA + CG (ours) | 188.44 | 5.03 |
| CogX-S (native) | 19.83 | — |
| CogX-S + CG (ours) | 206.32 | 10.41 |

These single-case measurements show that donor computation is the main optimization target on transfer hosts; the Wan measurements above quantify guidance-duration savings and the separate peak-memory comparison.

## H DENOISING DYNAMICS

We inspect four fixed videos for each task and seed (ONE/MIX, 42/43), giving sixteen videos in total. Instrumentation preserves their initial noise, final latents and exported video bytes. The two seeds provide separate replications of the within-trajectory measurements.

### H.1 MEASUREMENT AND NORMALIZATION

Correction strength measures the size of the exchange contribution relative to the live attention output after the same output projection. For a block group, video, update and receiver frame, we

compute

$$r = \frac{\mathrm{RMS}(\Delta o_{\mathcal{R}})}{\max\{\mathrm{RMS}(o_{\mathcal{Q}}), 10^{-8}\}}, \qquad \mathrm{RMS}(x) = \sqrt{|x|^{-1} \sum_j x_j^2}.$$

Here $\mathcal{R}$ contains receivers eligible for the selected exchange direction, whereas $\mathcal{Q}$ contains all live queries in that frame. The numerator and denominator pool their respective entries across the block group and channels. We average the resulting ratios over eligible frames, then equally over three ten-block groups and four videos. Early, middle and late summarize updates 0–9, 20–29 and 40–49. Larger values indicate a stronger internal correction; this statistic is distinct from detection accuracy and visual quality.

### H.2 Early, middle, and late corrections

Table 24: When is the CG (ours) exchange correction strongest? Entries average normalized correction magnitudes over four videos and three block groups for each task/seed. Early, middle and late cover updates 0–9, 20–29 and 40–49. Presence and absence refer to the two correction directions, not detector scores.

| Group | Correction | Early | Middle | Late |
|---|---|---|---|---|
| ONE 42 | Absence | 0.05572 | 0.01585 | 0.01132 |
| ONE 42 | Presence | 0.05854 | 0.01524 | 0.01238 |
| MIX 42 | Absence | 0.06103 | 0.01774 | 0.01092 |
| MIX 42 | Presence | 0.06828 | 0.02105 | 0.01438 |
| ONE 43 | Absence | 0.06457 | 0.01721 | 0.01002 |
| ONE 43 | Presence | 0.06507 | 0.01434 | 0.00922 |
| MIX 43 | Absence | 0.06792 | 0.01811 | 0.01374 |
| MIX 43 | Presence | 0.07855 | 0.01830 | 0.01346 |

**Early exchange corrections.** Both correction directions are strongest early in all four task/seed groups. The guidance-duration experiment tests whether this concentration permits shorter intervention without losing control. Exchange changes the values carried by attention, so a strong value correction need not change the attention weights. The additional diagnostics separate cross-state attention allocation from the propagation of a removed correction through later blocks.

## I Motion timing and action evidence

**Controlled comparison.** We generate Wan videos for twelve human-action prompts: clapping, kicking, punching, throwing, jumping, waving, bowing, squatting, raising both arms, stomping, swinging both arms and lunging. With seeds 42 and 43, we cross value exchange (off/on) with action-text gain $\gamma = 1$ or 1.5: $12 \times 2 \times 2 \times 2 = 96$ videos. Each row evaluates the same twelve prompts. Gain 1 leaves action-text strength unchanged; gain 1.5 amplifies it.

**TC motion-timing metric.** We use TC's strongest-motion-second statistic (Schiber et al., 2026). Its released evaluator uses Farnebäck optical flow (Farnebäck, 2003) on grayscale frames sampled at approximately 4 fps, averages whole-frame flow magnitudes within five one-second bins, and compares the largest bin with the requested second. Our 96-video ablation uses the same statistic; it is separate from the official 100-prompt movement benchmark. Camera motion or an unrelated action can yield a correct timing score because the statistic does not test action identity.

**Effect of value exchange.** At unit gain, exchange places peak motion in the requested window in two more videos out of twelve for each seed (+16.67 percentage points). MUSIQ changes by +0.41 and +0.08. Increasing text gain does not strengthen this effect: at gain 1.5, the counts are unchanged at seed 42 and decrease by two videos at seed 43. The comparison therefore distinguishes value exchange from simply increasing the strength of action text.

Table 25: CG (ours) motion ablation on twelve prompts per condition. Exchange off/on is crossed with action-text gain $\gamma$; all rows use the same prompts.

| Seed | Exchange / gain | Motion-window accuracy (%) | MUSIQ |
|---|---|---|---|
| 42 | Off / $\gamma = 1$ | 58.33 | 72.99 |
| 42 | On / $\gamma = 1$ | 75.00 | 73.40 |
| 42 | Off / $\gamma = 1.5$ | 66.67 | 71.53 |
| 42 | On / $\gamma = 1.5$ | 66.67 | 72.65 |
| 43 | Off / $\gamma = 1$ | 33.33 | 71.34 |
| 43 | On / $\gamma = 1$ | 50.00 | 71.42 |
| 43 | Off / $\gamma = 1.5$ | 58.33 | 71.03 |
| 43 | On / $\gamma = 1.5$ | 41.67 | 70.13 |

**Implementation checks.** All four batch/seed runs pass the implementation checks: gain 1 reproduces the unamplified reference exactly, prompt bindings remain fixed, and disabling the intervention restores the reference computation. The twelve-prompt comparison isolates the interaction between exchange and text gain; stronger action text does not consistently improve timing.

**Complementary requested-action evidence.** We test frozen InternVideo2 (Wang et al., 2024c) on eight fixed clips from this panel and forty controls: a static, camera-only, wrong-action, time-shifted and reversed version per clip. One-second windows, spaced by 0.5 seconds, each contain four frames resized to $224 \times 224$ pixels. Each temporal window is compared with the requested action and fixed descriptions of static poses, camera motion and unrelated actions. Relative support means that the requested-action cosine similarity exceeds every contrast in at least one window; it is not a calibrated presence probability. The requested action received support in 4/8 static and 5/8 camera-only controls, despite no action occurring. This violates the validation criterion that these controls receive no action support. The corresponding counts were 5/8 for unmodified clips, 0/8 for wrong-action queries, and 5/8 each for shifted and reversed clips. These are retrieval responses, not recognition accuracies: independent positive and wrong-action labels were not available. We therefore do not use this configuration to score action presence or action timing on the full panel. Its 96-video scores (792 additional window encodings) are retained as descriptive retrieval results. These controls distinguish semantic similarity from evidence that an action occurred at the requested time.

**Keypoint event spotting.** Because a video–text score responded to static poses, we also test explicit action events. YOLOv8[1] pose, person-segmentation and ball detectors run on every decoded frame; rules frozen before observation mark visible events: minimum hand separation for clapping, ankle separation for kicking, wrist excursions for waving, knee flexion for squatting, wrist elevation for raised arms, arm separation for swinging, ball–wrist separation after a held phase for throwing, the airborne apex for jumping and a foot stop after downward motion for stomping. Heights are compensated for background optical flow. When a rule's joints or ball are not visible, the clip is unobservable, not absent. Noise floors come only from static and camera-only controls. The initial rules produced three spurious events on 24 static and camera-only development controls, all from one camera-only squat clip. Eight of twelve development clips were unobservable because waist-up framing hid the hips. We retained the thresholds but used shoulder width when hips were out of frame and made paired-joint rules invariant to left/right label swaps.

We evaluated the revised rules on twelve held-out clips and sixty controls. They produced two spurious events on 24 static and camera-only controls, both from one camera-only raised-arms clip. Detected natural events shifted by exactly one second when one second was prepended, but tracking jitter under image translation still mimicked arm and knee extrema. Across the 96 generated videos, 50 had observable joints for their action rule. Table 26 reports event counts from these rules. Some camera-only clips still trigger detections, and the videos have no human-annotated event times, so these counts do not establish action-timing accuracy.

[1] https://github.com/ultralytics/ultralytics

Table 26: CG (ours) keypoint events by condition (twelve prompts each). Observable: clips whose rule joints are visible. In bin: an event inside the requested second. Multi: clips with more than one event. Flow: TC motion-window accuracy on all twelve clips. Counts are descriptive: camera-only controls still produced spurious events.

| Seed | Exchange / gain | Observable | In bin | In bin (%) | Multi | Flow (%) |
|---|---|---|---|---|---|---|
| 42 | Off / $\gamma = 1$ | 6 | 4 | 66.7 | 4 | 58.3 |
| 42 | On / $\gamma = 1$ | 6 | 2 | 33.3 | 3 | 75.0 |
| 42 | Off / $\gamma = 1.5$ | 6 | 5 | 83.3 | 4 | 66.7 |
| 42 | On / $\gamma = 1.5$ | 7 | 3 | 42.9 | 3 | 66.7 |
| 43 | Off / $\gamma = 1$ | 6 | 0 | 0.0 | 0 | 33.3 |
| 43 | On / $\gamma = 1$ | 6 | 1 | 16.7 | 3 | 50.0 |
| 43 | Off / $\gamma = 1.5$ | 6 | 0 | 0.0 | 0 | 58.3 |
| 43 | On / $\gamma = 1.5$ | 7 | 2 | 28.6 | 2 | 41.7 |

## J CogVideoX guidance calibration

The benchmark comparison uses two CogVideoX settings: CFG 6, suggested by the authors in their released inference example,[2] and CFG 3.5, selected for CG using development prompts. Including CFG 6 measures CG under the host authors' suggested guidance strength; CFG 3.5 measures the quality–timing balance used in our main comparison. Both retain the host's prediction space and sampler.

**Development comparison and choice.** We compare scalar CFG 3.5, 4, 4.5, 5 and 6 with two Haar variants on the same forty development prompts (Table 27). The selection rule seeks the highest MUSIQ while keeping TA, Presence and Absence within five points of both scalar reference settings (CFG 6 and 4.5) and improving MUSIQ over CFG 4.5. Haar triples denote coarse-active, coarse-inactive and detail gains.

CFG 3.5 satisfies the temporal bounds and gives the highest eligible MUSIQ: 54.65, improving on CFG 6 by 7.77 points. CFG 4 is excluded because Presence drops by 6.75 points relative to CFG 4.5; CFG 5 does not improve its MUSIQ. Haar 4.5/4.5/6 satisfies the bounds but has 2.29 lower MUSIQ than scalar 3.5. Haar 3.5/3.5/6 adds only 0.014 MUSIQ over scalar 3.5 and violates the Absence bound (8.50 points below CFG 6). CFG 4.5 and 6 offer lower MUSIQ, so scalar 3.5 is the main quality–timing choice. CFG 6 is retained as the authors' suggested setting in the benchmark comparison.

Table 27: CG (ours) guidance calibration on CogVideoX, forty prompts per setting. All scores are higher-is-better. Selection follows the temporal bounds and quality criterion described above.

| Setting | TA | Pres | Abs | MUSIQ | Selection reason |
|---|---|---|---|---|---|
| CFG 6 | 70.625 | 68.500 | 72.750 | 46.8820 | Authors' reference |
| CFG 4.5 | 69.875 | 69.000 | 70.750 | 51.5625 | Quality reference |
| CFG 5 | 67.750 | 66.250 | 69.250 | 51.2159 | MUSIQ below CFG 4.5 |
| CFG 4 | 65.750 | 62.250 | 69.250 | 53.5687 | Pres loss $> 5$ points |
| **CFG 3.5** | 67.250 | 66.000 | 68.500 | 54.6489 | **Selected** |
| Haar 4.5/4.5/6 | 69.750 | 68.000 | 71.500 | 52.3632 | MUSIQ below CFG 3.5 |
| Haar 3.5/3.5/6 | 68.625 | 73.000 | 64.250 | 54.6625 | Abs loss $> 5$ points |

**Benchmark results.** Both settings are evaluated on all eighty test prompts, with MUSIQ averaged over all 49 native frames per video. CFG 6 improves all three temporal metrics over published same-host TC; CFG 3.5 improves all four metrics, including MUSIQ (Table 28). The comparison shows that CG supports the authors' suggested guidance strength, while lower CFG provides a better balance of measured visual quality and temporal control. TC values are the published reference.

[2] https://huggingface.co/zai-org/CogVideoX-2b

Table 28: CogVideoX benchmark results, 80 prompts, published scoring (P). TC scores are published; bold: best per metric.

| Method / setting | TA | Pres | Abs | MUSIQ |
|---|---|---|---|---|
| CogX-S + TC | 63.75 | 47.62 | 78.41 | 46.67 |
| CogX-S + CG (ours), CFG 6 | **82.88** | **74.62** | **91.12** | 44.67 |
| CogX-S + CG (ours), CFG 3.5 | 75.06 | 69.50 | 80.63 | **49.53** |

# K Two-Object Composition Extension

## K.1 Motivation and scope

The two-object setting requires a continuing object $q_S$ to remain recognizable while a second object $q_T$ follows a temporal schedule. The core cross-state operator controls the content exchanged between different temporal states, but does not explicitly allocate spatial support to the two roles or modify same-state messages. We therefore augment it with spatial text permissions, regional retrieval, additional positive value replacement, and text-value amplification.

This extension is used uniformly in the reported two-object configuration. It is not used in the other evaluated settings. Its regions specify conditioning permissions rather than object segmentations or bounding boxes; they do not guarantee instance separation or constrain the complete extent of a rendered object.

## K.2 Notation and spatial permissions

We suppress the Transformer-block and attention-head indices. As in Sec. 4.1, $V$ denotes the live video-token values, while $V^{(0)}$ and $V^{(1)}$ denote values computed by the state-0 and state-1 donor branches. For the entrant schedule, define diagonal selectors

$$D_{ii} = c_{T,f(i)}, \qquad E = I - D. \tag{9}$$

Left multiplication selects receiving tokens; multiplication between an attention matrix and a value matrix selects source tokens.

Let $\mathcal{R}_S$ be a fixed central region of the spatial attention-token grid, repeated at every video time, and let $\mathcal{R}_T$ be its complement. Define

$$(\Pi_S)_{ii} = \mathbf{1}[i \in \mathcal{R}_S], \qquad \Pi_T = I - \Pi_S. \tag{10}$$

These spatial selectors are independent of the temporal schedule.

The additional components are active during the first half of sampling. With the descending inference index $\tau = N, \dots, 1$ from Sec. 3.1, their common activation function is

$$g_\tau = \mathbf{1}[\tau > \lfloor N/2 \rfloor]\,. \tag{11}$$

This sampling-step schedule is distinct from the video-time schedule $c_{T,f}$.

During this phase, the live and state-1 donor branches apply a spatial overlay to their existing text permissions: central queries cannot access the entrant's identity tokens $J_T$, and surround queries cannot access the continuing object's identity tokens $J_S$. The overlay only removes permissions; it never restores a token excluded by the underlying temporal-state mask. Consequently, live queries can access $J_T$ only in $\mathcal{R}_T$ at required temporal positions. The state-1 donor applies its spatial permissions at every temporal position.

The state-0 donor retains its original event-aware inactive permissions and receives no spatial overlay. The prompt is encoded once; all permissions act on the original contextual text features. The underlying state masks are shared by the configurations compared below, including the exclusion of the declared scene-change clause from the positive interpretation.

### K.3 REGIONAL RETRIEVAL

Let $A$ denote attention computed from the current branch's own queries and keys. For a query $i \in \mathcal{R}_T$, define restricted attention

$$R_{ij} = \frac{\mathbf{1}[j \in \mathcal{R}_T] \exp(a_{ij})}{\sum_{k \in \mathcal{R}_T} \exp(a_{ik})}, \qquad a_{ij} = \frac{Q_i K_j^\top}{\sqrt{d_h}}, \tag{12}$$

where the queries and keys include the host model's native normalization and positional processing. The source set includes valid tokens in $\mathcal{R}_T$ across all video times, rather than only the receiving token's frame.

We use

$$B_{ij} = \begin{cases} (1 - g_\tau \lambda) A_{ij} + g_\tau \lambda R_{ij}, & i \in \mathcal{R}_T, \\ A_{ij}, & i \in \mathcal{R}_S, \end{cases} \qquad \lambda = 0.5. \tag{13}$$

Thus, central queries retain global retrieval, while surround queries retain a global component and gain a region-restricted component. This does not disconnect the two regions.

The live branch, state-1 donor, and unconditional CFG branch each use Eq. (13), computed from their own queries and keys. The state-0 donor retains native self-attention. The unconditional branch receives neither object-specific text permissions nor donor-value replacements.

### K.4 REGIONAL POSITIVE VALUE REPLACEMENT

Using $B$ in place of $A$, the core cross-state operator is

$$Y^{\text{core}}(B) = BV + EBD\big(V^{(0)} - V\big) + DBE\big(V^{(1)} - V\big). \tag{14}$$

The second term supplies state-0 values to inactive receivers from required source positions. The third supplies state-1 values to required receivers from inactive source positions.

The composition extension additionally uses

$$\begin{aligned} Y^{\text{comp}} = {} & Y^{\text{core}}(B) \\ & + g_\tau D \Pi_T B D \Pi_T \big(V^{(1)} - V\big). \end{aligned} \tag{15}$$

The added term replaces live values with positive-donor values on messages between two required tokens in the entrant region. These source positions are disjoint from those of the original positive cross-state term, whose sources lie in $E$. The original cross-state replacements remain intact.

Despite its additive residual form, this term performs value replacement: on the selected messages, the contribution changes from $B_{ij} V_j$ to $B_{ij} V_j^{(1)}$. It does not add a second copy of the donor contribution. The combined attention output passes through the native output projection and residual modulation once.

Unlike the core operator alone, this configuration changes both attention allocation and a subset of same-state values. Its positive donor is also computed using regional retrieval and spatial text permissions. The core operator's preservation properties therefore do not apply to the complete extension.

### K.5 TEXT-VALUE SUPPORT FOR BOTH OBJECTS

To strengthen the two object identities within their permitted regions, we amplify their text-value contributions in the live branch only. Let $\alpha_{ij}$ be the live text-attention weights after the temporal and spatial permissions have been applied, and let $V_j^{\text{text}}$ and $b_V$ be the text-value vector and projection

bias, respectively. Define $s_i = (\Pi_S)_{ii}$ and $h_i = (\Pi_T)_{ii}$. The modified output is

$$\begin{aligned}\widetilde{Z}_i^{\text{text}} = \sum_j \alpha_{ij} V_j^{\text{text}} \\ + g_\tau(\gamma_{\text{obj}} - 1)\Bigg[ s_i \sum_{j \in J_S} \alpha_{ij}\big(V_j^{\text{text}} - b_V\big) \\ + c_{T,f(i)} h_i \sum_{j \in J_T} \alpha_{ij}\big(V_j^{\text{text}} - b_V\big)\Bigg],\end{aligned} \tag{16}$$

with $\gamma_{\text{obj}} = 1.5$.

The continuing object receives support in its permitted region at every video time; the entrant receives support in its permitted region only when required. This operation changes text values, not attention logits or their softmax normalization. Subtracting $b_V$ prevents amplification of the shared projection bias. The donor and unconditional branches receive no text-value gain.

This is the same bias-preserving form used for action support in Eq. (7), but with different token sets, query support, and branch scope. All non-object text values remain unchanged.

### K.6 Branch policies and implementation

Table 29 summarizes the additional operations during the active sampling phase. "Spatial overlay" means the permission restrictions defined above; the underlying state-dependent text permissions remain in effect throughout.

Table 29: Branch policies for the CG (ours) two-object extension during the first half of sampling. Each branch computes attention using its own features.

| Branch | Spatial overlay | Retrieval | Value exchange | Text gain |
|---|---|---|---|---|
| Live | Yes | Mixed | Eq. (15) | S+T |
| State-1 donor | Yes | Mixed | None | None |
| State-0 donor | No | Native | None | None |
| Unconditional | No | Mixed | None | None |

For Wan, the attention-token grid has spatial dimensions $H_a = 30$ and $W_a = 52$. Using zero-based coordinates, the central region is

$$\mathcal{R}_S = \{i : a_H \le r_i < H_a - a_H,\ a_W \le u_i < W_a - a_W\}, \tag{17}$$

where $a_H = \lfloor H_a/4 \rfloor$ and $a_W = \lfloor W_a/4 \rfloor$. Thus, the central region covers rows $[7, 23)$ and columns $[13, 39)$: 416 spatial tokens per frame, with 1,144 in the surrounding entrant region. The same geometry is used at every video time and for every example.

The additions operate in all 30 Transformer blocks during the first 25 of 50 sampling updates. They are then disabled together: text access returns to the underlying temporal-state masks, retrieval becomes native, the extra same-state replacement is removed, and the object text gain is disabled. Subsequent computation therefore uses the core configuration at the current generative state; it does not revert the trajectory to one generated without the extension.

Restricted attention is evaluated directly on gathered region tokens after native query/key normalization and positional encoding, preserving their original positions. It is not recovered by subtracting rounded global attention outputs and dividing by a regional mass. The implementation retains global attention windows and excludes padded source positions.

For the query-dependent text gain, the implementation evaluates two cross-attention outputs with identical queries, keys, and permissions: one amplifies $J_S$, and the other amplifies both $J_S$ and $J_T$. Required queries select the second output; inactive queries select the first. Each query receives one output, not their sum. This introduces additional cross-attention work but no additional donor branch or sampling trajectory.

### K.7 Component comparison and limitations

Table 30 compares the core configuration, the regional extension without object text amplification, and the complete extension. All rows use the same 82-prompt evaluation panel, seed, noise bank, batch composition, and evaluator. The core configuration includes the same event-aware state permissions used by the extensions.

Table 30: CG (ours) component comparison on the 82-prompt two-object panel. Temporal metrics are percentages. The middle row retains spatial permissions, mixed retrieval, and regional positive replacement, but omits Eq. (16).

| Configuration | TA | Presence | Absence | MUSIQ |
|---|---|---|---|---|
| Core + event-aware masks | 57.26 | 42.24 | 73.78 | 70.99 |
| Regional composition, no text gain | 61.90 | 57.54 | 66.71 | 71.32 |
| Regional composition + S+T text gain | 67.83 | 60.42 | 75.98 | 70.89 |

Relative to the regional configuration without text gain, S+T support increases correct requested-present frame decisions from 519/902 (57.54%) to 545/902 (60.42%), and requested-absent decisions from 547/820 (66.71%) to 623/820 (75.98%). MUSIQ decreases slightly from 71.32 to 70.89. The core-to-regional comparison changes multiple regional components jointly; the regional-to-full comparison isolates the joint S+T text-value gain, but not the contribution of each object separately. These results support the complete configuration for this two-object setting, rather than establishing a guarantee of instance preservation or correct timing. Spatial permissions can encourage peripheral or clipped entrants, donor values need not represent clean object-specific content, and some examples regress. Because video self-attention couples temporal positions, restricting a direct correction to required queries does not guarantee that inactive output frames remain unchanged.

## L Mixed-lifecycle prompt construction

**Object vocabulary and prompt scheme.** We use only the 80 COCO object classes recognized by the YOLO evaluator, with one prompt per class. The benchmark extends TempoControl's explicitly timed, object-centered prompt templates (Schiber et al., 2026) to mixed lifecycles. This follows the object-based evaluation setting used by TC, whose two-object benchmark draws its object pairs from VBench (Huang et al., 2024); the exit and return clauses below are our extension.

**Manually chosen event wording.** For each class, we decide whether the object can naturally leave the scene and manually choose an appropriate exit phrase and, when required, a return phrase. Animals can walk, run or fly away; vehicles can drive, roll or sail away. Objects that cannot move independently are carried away, or leave view through a camera pan when that is more natural. Return wording uses the corresponding action and specifies the same object. Table 31 lists the resulting phrase families. This controls visibility through a plausible event rather than requiring every object to vanish in place.

**Random time intervals and shared evaluation.** After choosing the wording, we randomly select the requested time periods for the lifecycle events. The final set contains 27 entry–exit, 27 exit, and 26 exit–return schedules. We insert these times into the prompt and derive its binary presence schedule from the same intervals. We then fix the resulting 80 prompt–schedule pairs and use them unchanged for every method. The exact prompts, object labels and schedules are provided in `data/test/mixed.csv`; the complete prompt list follows below.

### L.1 Complete mixed-lifecycle prompts

The following eighty prompts appear in the order of the released registry. Bold labels identify the controlled object; the following text is the exact generation prompt. All methods receive the same prompt and corresponding binary schedule from `data/test/mixed.csv`.

Table 31: Fixed mixed-prompt event phrases. Exit and return deadlines are derived from the requested schedule.

| Family | Exit | Return |
|---|---|---|
| Fly | flies away | flies back into frame |
| Run | runs away | runs back into frame |
| Walk | walks away | walks back into frame |
| Drive | drives away | drives back into frame |
| Roll | rolls away | rolls back into frame |
| Sail | sails away | sails back into frame |
| Depart | pulls away | pulls back into frame |
| Carried | is carried away | is carried back into frame |
| Camera | is left behind as the camera pans away | re-enters the frame as the camera pans back |

1. **airplane.** An empty scene. Suddenly, during the third second of the video, an airplane appears out of nowhere, drawing all attention. The airplane flies away and is completely out of view by the start of the fifth second of the video.
2. **apple.** The video begins with a serene view centered on the apple. The apple is carried away and is completely out of view by the start of the third second of the video.
3. **backpack.** The video begins with a serene view centered on the backpack. The backpack is carried away and is completely out of view by the start of the fourth second of the video. The same backpack is carried back into frame and is fully visible again by the start of the fifth second of the video.
4. **banana.** An empty scene. Suddenly, during the second second of the video, a banana appears out of nowhere, drawing all attention. The banana is carried away and is completely out of view by the start of the fourth second of the video.
5. **baseball bat.** An empty scene. Suddenly, during the third second of the video, a baseball bat appears out of nowhere, drawing all attention. The baseball bat is carried away and is completely out of view by the start of the fifth second of the video.
6. **baseball glove.** The video begins with a serene view centered on the baseball glove. The baseball glove is carried away and is completely out of view by the start of the third second of the video.
7. **bear.** The video begins with a serene view centered on the bear. The bear runs away and is completely out of view by the start of the fourth second of the video. The same bear runs back into frame and is fully visible again by the start of the fifth second of the video.
8. **bed.** An empty scene. Suddenly, during the second second of the video, a bed appears out of nowhere, drawing all attention. The bed is left behind as the camera pans away and is completely out of view by the start of the fourth second of the video.
9. **bench.** The video begins with a serene view centered on the bench. The bench is left behind as the camera pans away and is completely out of view by the start of the third second of the video.
10. **bicycle.** The video begins with a serene view centered on the bicycle. The bicycle rolls away and is completely out of view by the start of the third second of the video.
11. **bird.** The video begins with a serene view centered on the bird. The bird flies away and is completely out of view by the start of the third second of the video.
12. **boat.** An empty scene. Suddenly, during the second second of the video, a boat appears out of nowhere, drawing all attention. The boat sails away and is completely out of view by the start of the fourth second of the video.
13. **book.** The video begins with a serene view centered on the book. The book is carried away and is completely out of view by the start of the fourth second of the video. The same book is carried back into frame and is fully visible again by the start of the fifth second of the video.

14. **bottle.** An empty scene. Suddenly, during the second second of the video, a bottle appears out of nowhere, drawing all attention. The bottle is carried away and is completely out of view by the start of the fourth second of the video.
15. **bowl.** An empty scene. Suddenly, during the second second of the video, a bowl appears out of nowhere, drawing all attention. The bowl is carried away and is completely out of view by the start of the fifth second of the video.
16. **broccoli.** The video begins with a serene view centered on the broccoli. The broccoli is carried away and is completely out of view by the start of the third second of the video.
17. **bus.** The video begins with a serene view centered on the bus. The bus drives away and is completely out of view by the start of the third second of the video.
18. **cake.** The video begins with a serene view centered on the cake. The cake is carried away and is completely out of view by the start of the fourth second of the video. The same cake is carried back into frame and is fully visible again by the start of the fifth second of the video.
19. **car.** An empty scene. Suddenly, during the third second of the video, a car appears out of nowhere, drawing all attention. The car drives away and is completely out of view by the start of the fifth second of the video.
20. **carrot.** The video begins with a serene view centered on the carrot. The carrot is carried away and is completely out of view by the start of the fourth second of the video. The same carrot is carried back into frame and is fully visible again by the start of the fifth second of the video.
21. **cat.** The video begins with a serene view centered on the cat. The cat runs away and is completely out of view by the start of the third second of the video.
22. **cell phone.** An empty scene. Suddenly, during the second second of the video, a cell phone appears out of nowhere, drawing all attention. The cell phone is carried away and is completely out of view by the start of the fourth second of the video.
23. **chair.** The video begins with a serene view centered on the chair. The chair is carried away and is completely out of view by the start of the third second of the video.
24. **clock.** The video begins with a serene view centered on the clock. The clock is left behind as the camera pans away and is completely out of view by the start of the fourth second of the video. The same clock re-enters the frame as the camera pans back and is fully visible again by the start of the fifth second of the video.
25. **couch.** An empty scene. Suddenly, during the second second of the video, a couch appears out of nowhere, drawing all attention. The couch is left behind as the camera pans away and is completely out of view by the start of the fourth second of the video.
26. **cow.** An empty scene. Suddenly, during the second second of the video, a cow appears out of nowhere, drawing all attention. The cow walks away and is completely out of view by the start of the fifth second of the video.
27. **cup.** The video begins with a serene view centered on the cup. The cup is carried away and is completely out of view by the start of the fourth second of the video. The same cup is carried back into frame and is fully visible again by the start of the fifth second of the video.
28. **dining table.** An empty scene. Suddenly, during the third second of the video, a dining table appears out of nowhere, drawing all attention. The dining table is left behind as the camera pans away and is completely out of view by the start of the fifth second of the video.
29. **dog.** An empty scene. Suddenly, during the second second of the video, a dog appears out of nowhere, drawing all attention. The dog runs away and is completely out of view by the start of the fifth second of the video.
30. **donut.** An empty scene. Suddenly, during the second second of the video, a donut appears out of nowhere, drawing all attention. The donut is carried away and is completely out of view by the start of the fourth second of the video.
31. **elephant.** The video begins with a serene view centered on the elephant. The elephant walks away and is completely out of view by the start of the third second of the video.

32. **fire hydrant.** The video begins with a serene view centered on the fire hydrant. The fire hydrant is left behind as the camera pans away and is completely out of view by the start of the third second of the video.
33. **fork.** An empty scene. Suddenly, during the third second of the video, a fork appears out of nowhere, drawing all attention. The fork is carried away and is completely out of view by the start of the fifth second of the video.
34. **frisbee.** The video begins with a serene view centered on the frisbee. The frisbee flies away and is completely out of view by the start of the fourth second of the video. The same frisbee flies back into frame and is fully visible again by the start of the fifth second of the video.
35. **giraffe.** The video begins with a serene view centered on the giraffe. The giraffe walks away and is completely out of view by the start of the fourth second of the video. The same giraffe walks back into frame and is fully visible again by the start of the fifth second of the video.
36. **hair drier.** The video begins with a serene view centered on the hair drier. The hair drier is carried away and is completely out of view by the start of the third second of the video.
37. **handbag.** The video begins with a serene view centered on the handbag. The handbag is carried away and is completely out of view by the start of the fourth second of the video. The same handbag is carried back into frame and is fully visible again by the start of the fifth second of the video.
38. **horse.** The video begins with a serene view centered on the horse. The horse runs away and is completely out of view by the start of the third second of the video.
39. **hot dog.** The video begins with a serene view centered on the hot dog. The hot dog is carried away and is completely out of view by the start of the third second of the video.
40. **keyboard.** An empty scene. Suddenly, during the second second of the video, a keyboard appears out of nowhere, drawing all attention. The keyboard is carried away and is completely out of view by the start of the fifth second of the video.
41. **kite.** The video begins with a serene view centered on the kite. The kite flies away and is completely out of view by the start of the third second of the video.
42. **knife.** The video begins with a serene view centered on the knife. The knife is carried away and is completely out of view by the start of the third second of the video.
43. **laptop.** The video begins with a serene view centered on the laptop. The laptop is carried away and is completely out of view by the start of the fourth second of the video. The same laptop is carried back into frame and is fully visible again by the start of the fifth second of the video.
44. **microwave.** The video begins with a serene view centered on the microwave. The microwave is left behind as the camera pans away and is completely out of view by the start of the fourth second of the video. The same microwave re-enters the frame as the camera pans back and is fully visible again by the start of the fifth second of the video.
45. **motorcycle.** The video begins with a serene view centered on the motorcycle. The motorcycle drives away and is completely out of view by the start of the third second of the video.
46. **mouse.** The video begins with a serene view centered on the mouse. The mouse is carried away and is completely out of view by the start of the fourth second of the video. The same mouse is carried back into frame and is fully visible again by the start of the fifth second of the video.
47. **orange.** An empty scene. Suddenly, during the second second of the video, an orange appears out of nowhere, drawing all attention. The orange is carried away and is completely out of view by the start of the fourth second of the video.
48. **oven.** The video begins with a serene view centered on the oven. The oven is left behind as the camera pans away and is completely out of view by the start of the third second of the video.

49. **parking meter.** An empty scene. Suddenly, during the second second of the video, a parking meter appears out of nowhere, drawing all attention. The parking meter is left behind as the camera pans away and is completely out of view by the start of the fifth second of the video.
50. **person.** An empty scene. Suddenly, during the second second of the video, a person appears out of nowhere, drawing all attention. The person walks away and is completely out of view by the start of the fourth second of the video.
51. **pizza.** The video begins with a serene view centered on the pizza. The pizza is carried away and is completely out of view by the start of the third second of the video.
52. **potted plant.** The video begins with a serene view centered on the potted plant. The potted plant is carried away and is completely out of view by the start of the fourth second of the video. The same potted plant is carried back into frame and is fully visible again by the start of the fifth second of the video.
53. **refrigerator.** The video begins with a serene view centered on the refrigerator. The refrigerator is left behind as the camera pans away and is completely out of view by the start of the third second of the video.
54. **remote.** The video begins with a serene view centered on the remote. The remote is carried away and is completely out of view by the start of the fourth second of the video. The same remote is carried back into frame and is fully visible again by the start of the fifth second of the video.
55. **sandwich.** An empty scene. Suddenly, during the third second of the video, a sandwich appears out of nowhere, drawing all attention. The sandwich is carried away and is completely out of view by the start of the fifth second of the video.
56. **scissors.** The video begins with a serene view centered on the scissors. The scissors are carried away and are completely out of view by the start of the third second of the video.
57. **sheep.** The video begins with a serene view centered on the sheep. The sheep runs away and is completely out of view by the start of the fourth second of the video. The same sheep runs back into frame and is fully visible again by the start of the fifth second of the video.
58. **sink.** The video begins with a serene view centered on the sink. The sink is left behind as the camera pans away and is completely out of view by the start of the fourth second of the video. The same sink re-enters the frame as the camera pans back and is fully visible again by the start of the fifth second of the video.
59. **skateboard.** The video begins with a serene view centered on the skateboard. The skateboard rolls away and is completely out of view by the start of the fourth second of the video. The same skateboard rolls back into frame and is fully visible again by the start of the fifth second of the video.
60. **skis.** The video begins with a serene view centered on the skis. The skis are carried away and are completely out of view by the start of the fourth second of the video. The same skis are carried back into frame and are fully visible again by the start of the fifth second of the video.
61. **snowboard.** The video begins with a serene view centered on the snowboard. The snowboard is carried away and is completely out of view by the start of the fourth second of the video. The same snowboard is carried back into frame and is fully visible again by the start of the fifth second of the video.
62. **spoon.** An empty scene. Suddenly, during the second second of the video, a spoon appears out of nowhere, drawing all attention. The spoon is carried away and is completely out of view by the start of the fifth second of the video.
63. **sports ball.** The video begins with a serene view centered on the sports ball. The sports ball rolls away and is completely out of view by the start of the third second of the video.
64. **stop sign.** The video begins with a serene view centered on the stop sign. The stop sign is left behind as the camera pans away and is completely out of view by the start of the third second of the video.
65. **suitcase.** The video begins with a serene view centered on the suitcase. The suitcase is carried away and is completely out of view by the start of the third second of the video.

66. **surfboard.** The video begins with a serene view centered on the surfboard. The surfboard is carried away and is completely out of view by the start of the fourth second of the video. The same surfboard is carried back into frame and is fully visible again by the start of the fifth second of the video.
67. **teddy bear.** An empty scene. Suddenly, during the second second of the video, a teddy bear appears out of nowhere, drawing all attention. The teddy bear is carried away and is completely out of view by the start of the fourth second of the video.
68. **tennis racket.** An empty scene. Suddenly, during the second second of the video, a tennis racket appears out of nowhere, drawing all attention. The tennis racket is carried away and is completely out of view by the start of the fourth second of the video.
69. **tie.** The video begins with a serene view centered on the tie. The tie is carried away and is completely out of view by the start of the third second of the video.
70. **toaster.** The video begins with a serene view centered on the toaster. The toaster is carried away and is completely out of view by the start of the fourth second of the video. The same toaster is carried back into frame and is fully visible again by the start of the fifth second of the video.
71. **toilet.** An empty scene. Suddenly, during the second second of the video, a toilet appears out of nowhere, drawing all attention. The toilet is left behind as the camera pans away and is completely out of view by the start of the fourth second of the video.
72. **toothbrush.** An empty scene. Suddenly, during the second second of the video, a toothbrush appears out of nowhere, drawing all attention. The toothbrush is carried away and is completely out of view by the start of the fourth second of the video.
73. **traffic light.** The video begins with a serene view centered on the traffic light. The traffic light is left behind as the camera pans away and is completely out of view by the start of the third second of the video.
74. **train.** The video begins with a serene view centered on the train. The train pulls away and is completely out of view by the start of the fourth second of the video. The same train pulls back into frame and is fully visible again by the start of the fifth second of the video.
75. **truck.** The video begins with a serene view centered on the truck. The truck drives away and is completely out of view by the start of the fourth second of the video. The same truck drives back into frame and is fully visible again by the start of the fifth second of the video.
76. **tv.** The video begins with a serene view centered on the tv. The tv is left behind as the camera pans away and is completely out of view by the start of the fourth second of the video. The same tv re-enters the frame as the camera pans back and is fully visible again by the start of the fifth second of the video.
77. **umbrella.** An empty scene. Suddenly, during the third second of the video, an umbrella appears out of nowhere, drawing all attention. The umbrella is carried away and is completely out of view by the start of the fifth second of the video.
78. **vase.** The video begins with a serene view centered on the vase. The vase is carried away and is completely out of view by the start of the third second of the video.
79. **wine glass.** The video begins with a serene view centered on the wine glass. The wine glass is carried away and is completely out of view by the start of the fourth second of the video. The same wine glass is carried back into frame and is fully visible again by the start of the fifth second of the video.
80. **zebra.** An empty scene. Suddenly, during the second second of the video, a zebra appears out of nowhere, drawing all attention. The zebra runs away and is completely out of view by the start of the fourth second of the video.

# M QUALITATIVE COMPARISONS

These examples illustrate appearance and disappearance control. Methods receive the same timed prompt and are shown at identical timestamps and image scale. Captions describe the visible transition; aggregate performance is reported in the benchmark tables.

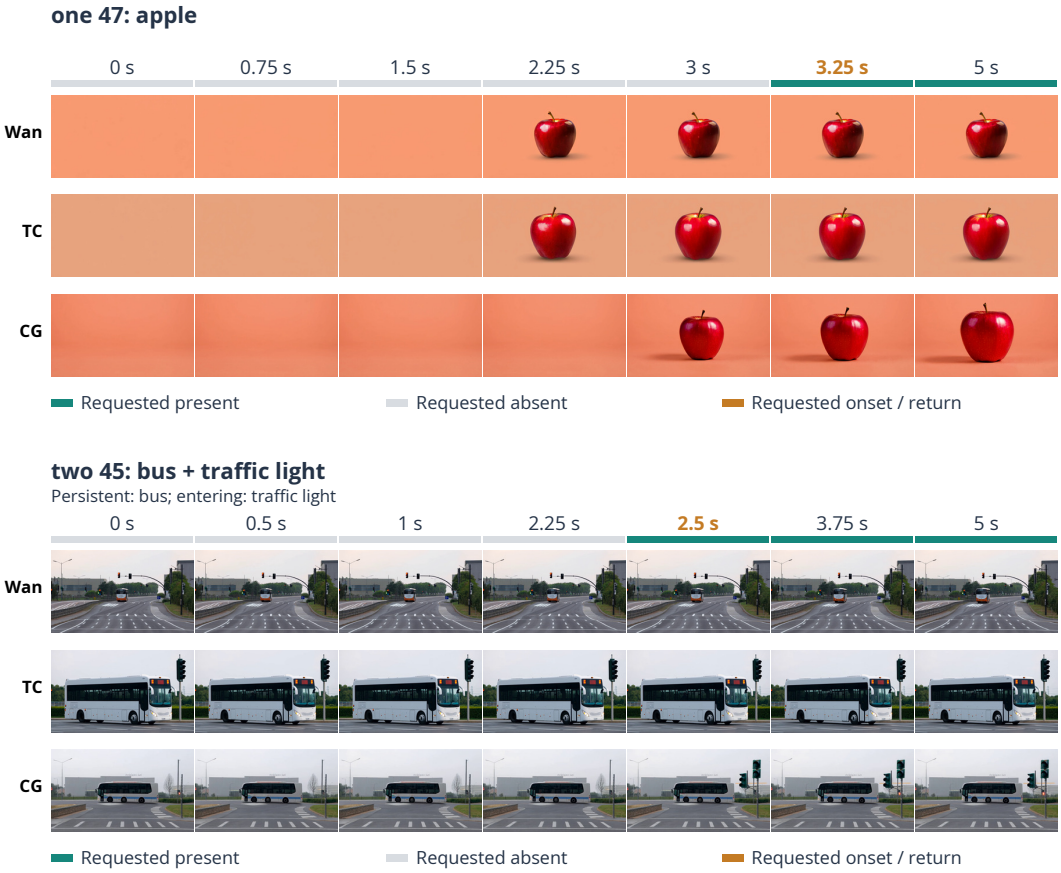



Figure 18: Apple: CG delays appearance toward the requested onset, while remaining early. Traffic light: CG delays the entering object while the bus remains visible.

Figures 29–33 compare Wan, TC and CG on five movement prompts using the website videos. Each figure follows the full five-second clip in three consecutive strips, at 0.25 s intervals. Synchronized columns reveal both action timing and behavior outside the requested interval.

## M.1 TRANSFER TO CONTINUOUS AND DISCRETE HOSTS

Figures 34 and 35 show two examples per host from the headline configurations: selected scalar CFG 3.5 for CogVideoX and the evaluated URSA configuration. Each pair shares the timed prompt, seed 42, and recorded initial latent. Selection uses the existing 80-prompt results per host, followed by visual review of all 49 frames in each selected video to assess first visible appearance. The prompt template is “An empty scene. Suddenly, at [time] seconds into the video, a [object] appears out of nowhere, drawing all attention”. The title gives the object and requested onset. Columns show selected snapshots at their actual timestamps, not a uniformly spaced time axis. Gray/green bars indicate requested absence/presence; all method names are black.

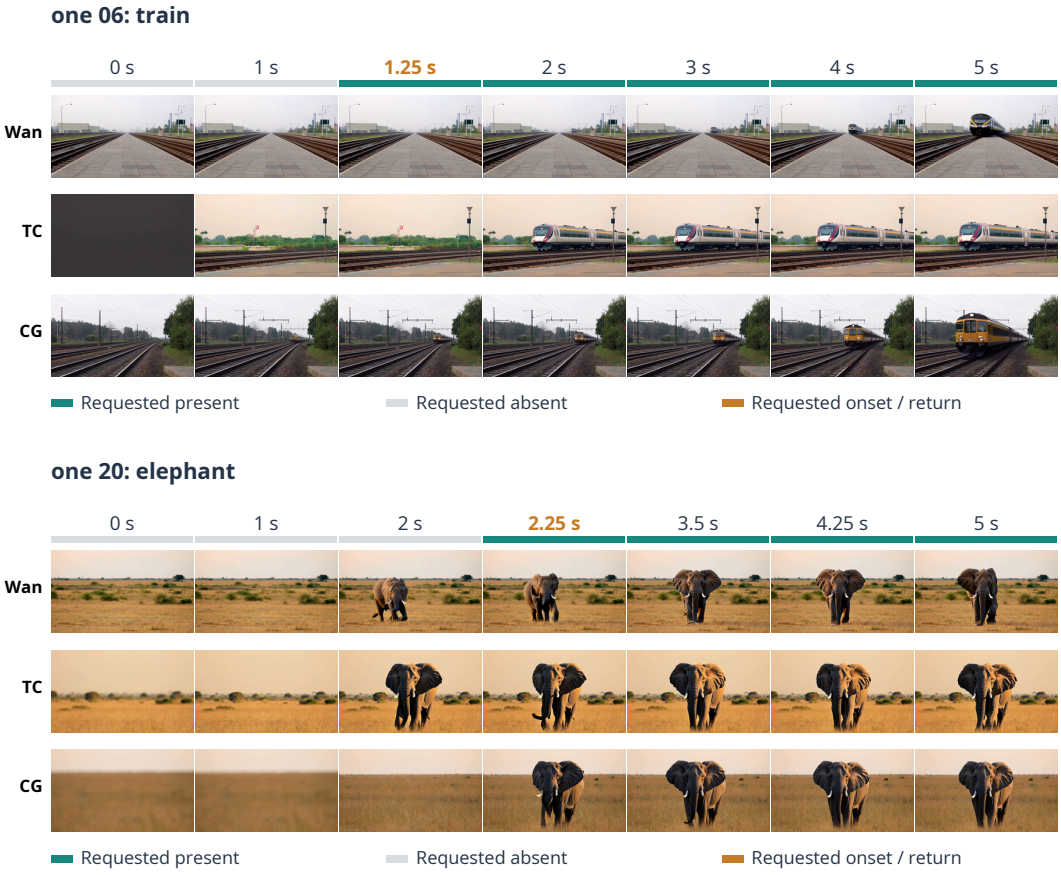



Figure 19: Train: both guided methods place appearance closer to the requested onset than the late native train. Elephant appearance. CG delays the elephant toward the requested onset at 2.25s; Wan and TC introduce it earlier. The elephant remains visible thereafter.

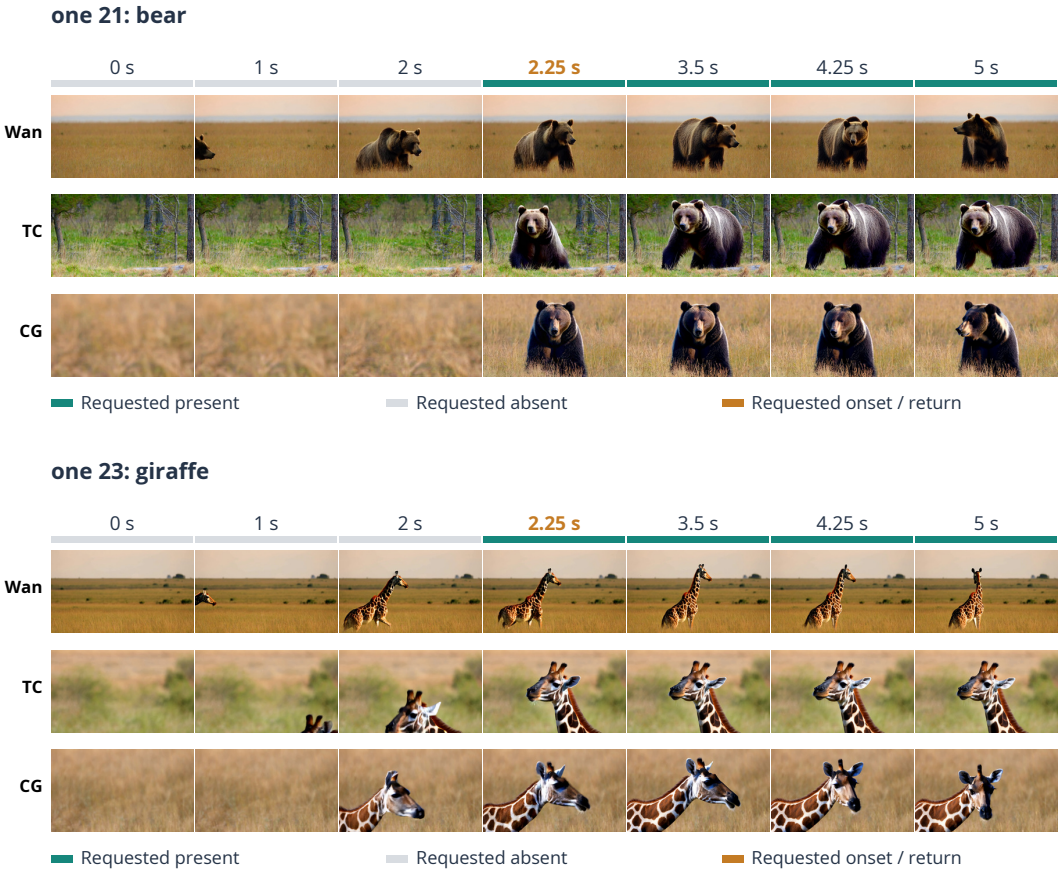



Figure 20: Bear appearance. Both guided videos delay the bear toward the requested onset at 2.25s relative to Wan. Giraffe appearance. CG delays entry relative to Wan and TC, bringing it closer to the requested onset at 2.25s, although its appearance is still early.

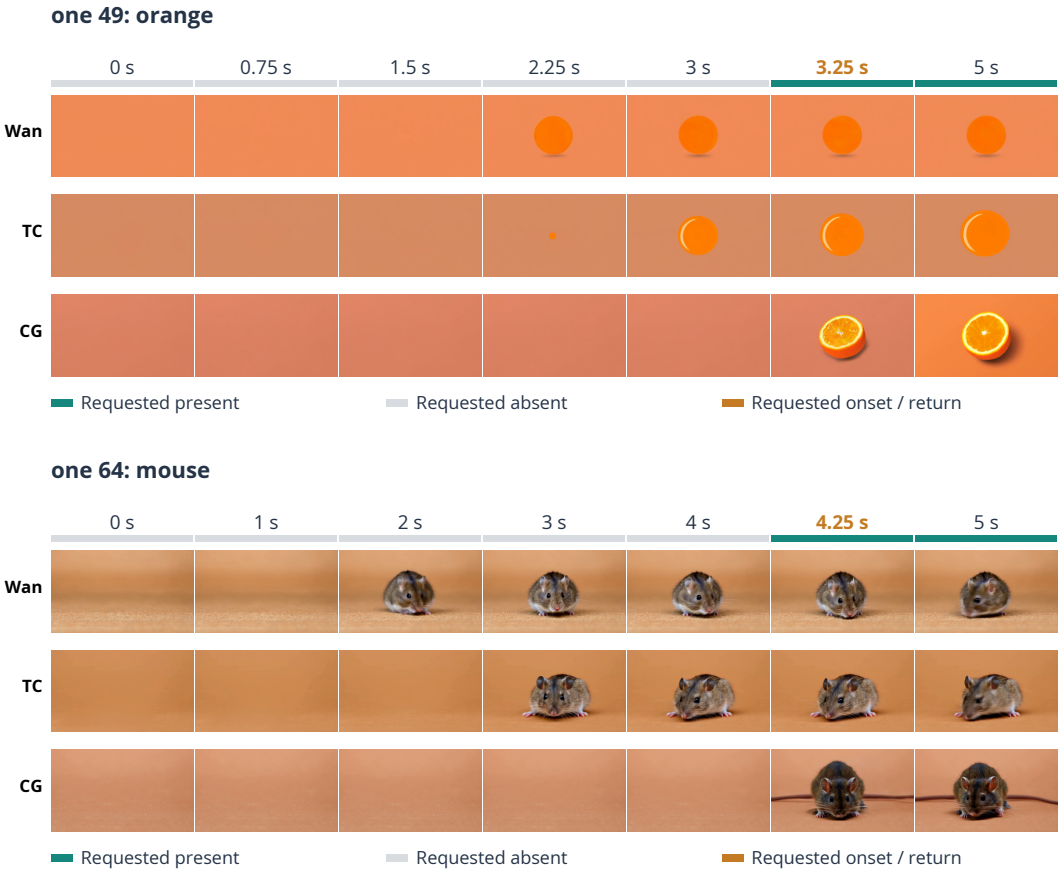



Figure 21: Orange: CG delays appearance to the requested interval while both comparators introduce the object early. Mouse: CG delays appearance toward the requested late onset and keeps the object visible afterward.

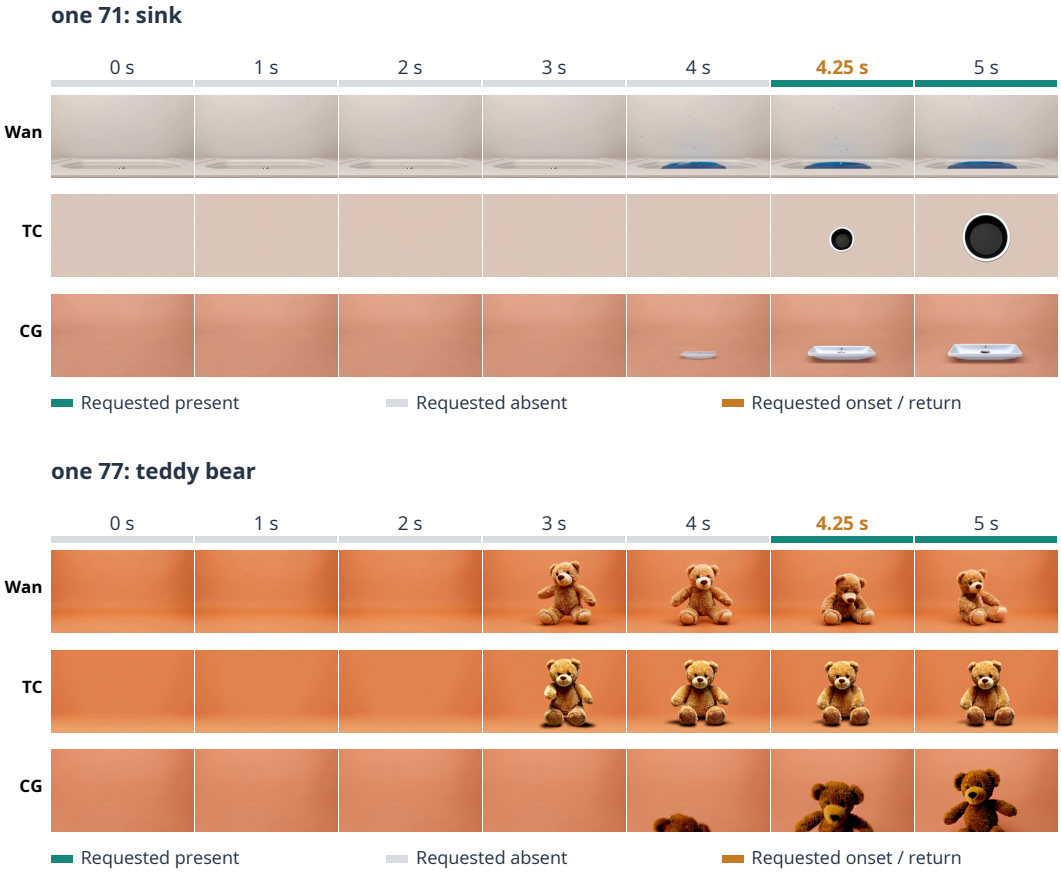



Figure 22: Sink: CG forms a recognizable basin near the requested late onset, while Wan shows the fixture from the first frame. Teddy bear: CG better respects the requested late entrance, avoiding the earlier appearances in both Wan and TC.

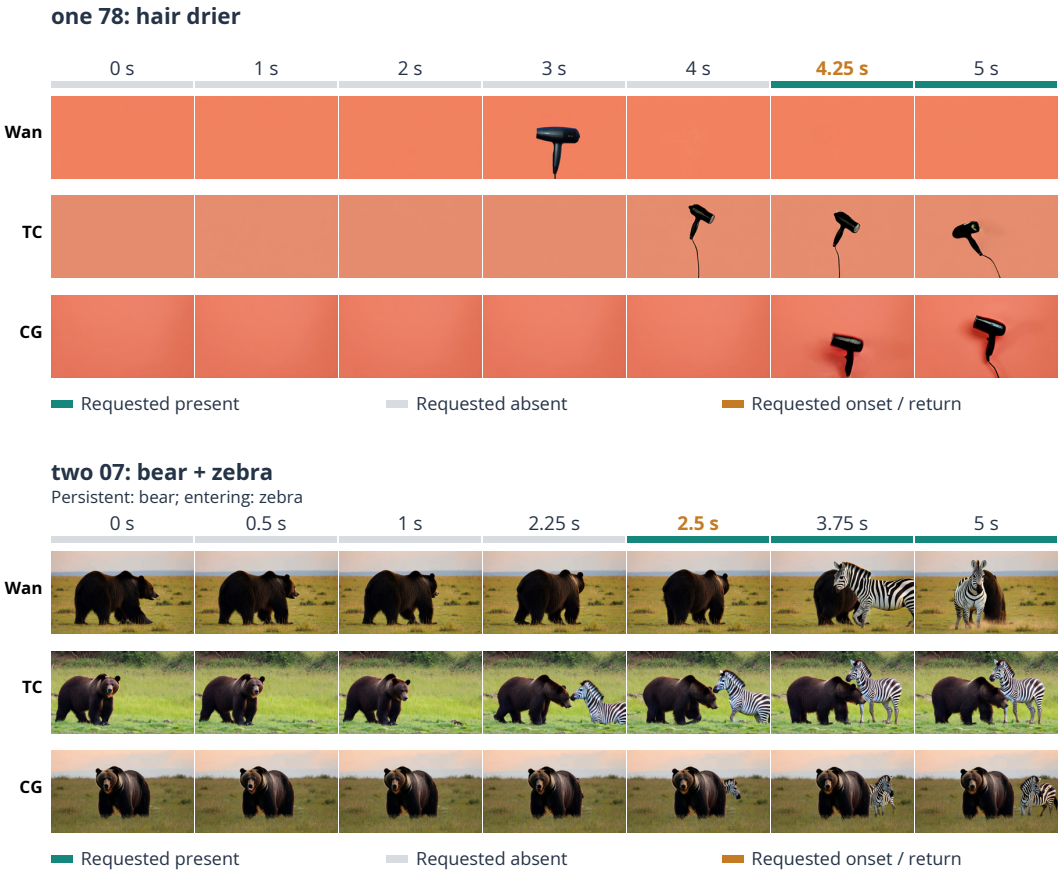



Figure 23: Hair drier: CG delays appearance toward the requested late onset and retains the object afterward. Bear and zebra: CG introduces the zebra near the requested second half while retaining the bear.




Figure 24: Suitcase and vase: CG introduces the vase closer to the requested second-half onset. Boat and airplane: CG delays the distant airplane until the requested second half; both comparators show an airplane from the start.

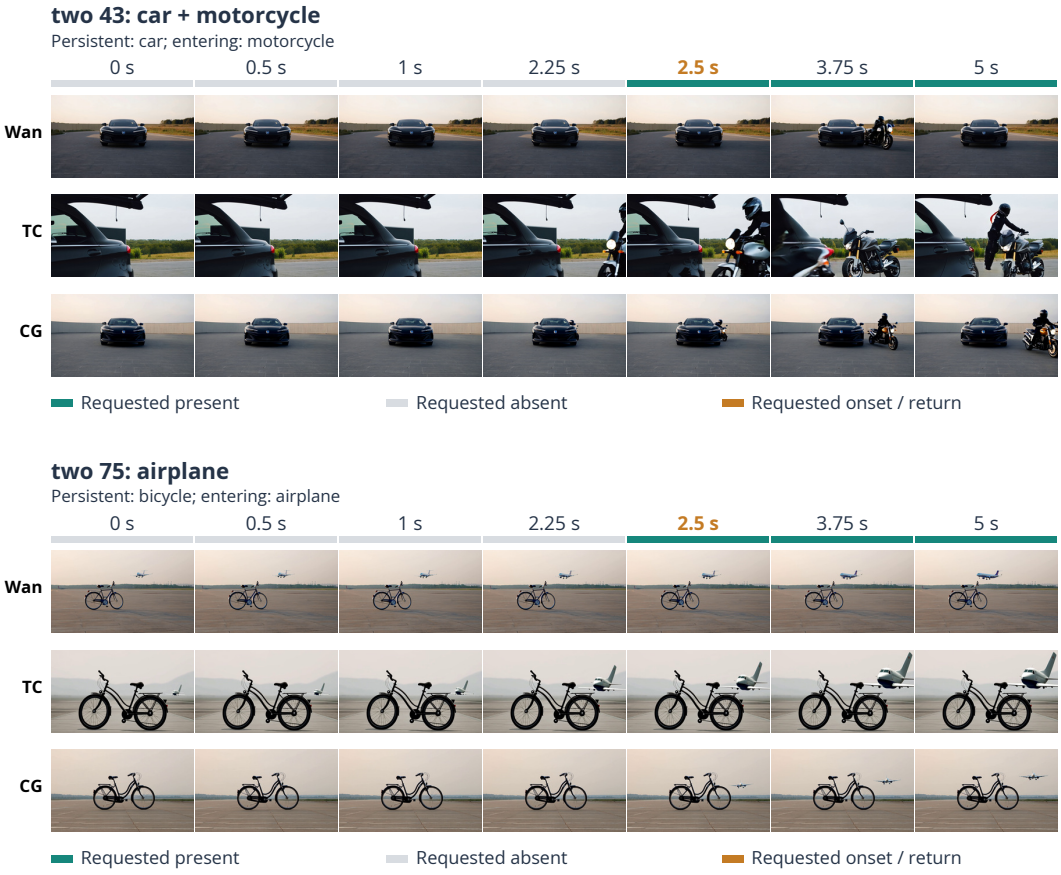



Figure 25: Car and motorcycle: CG introduces the motorcycle near the requested second half and retains both objects. Bicycle and airplane: CG introduces the airplane near the requested onset while keeping the bicycle visible; both comparators show the airplane from the start.

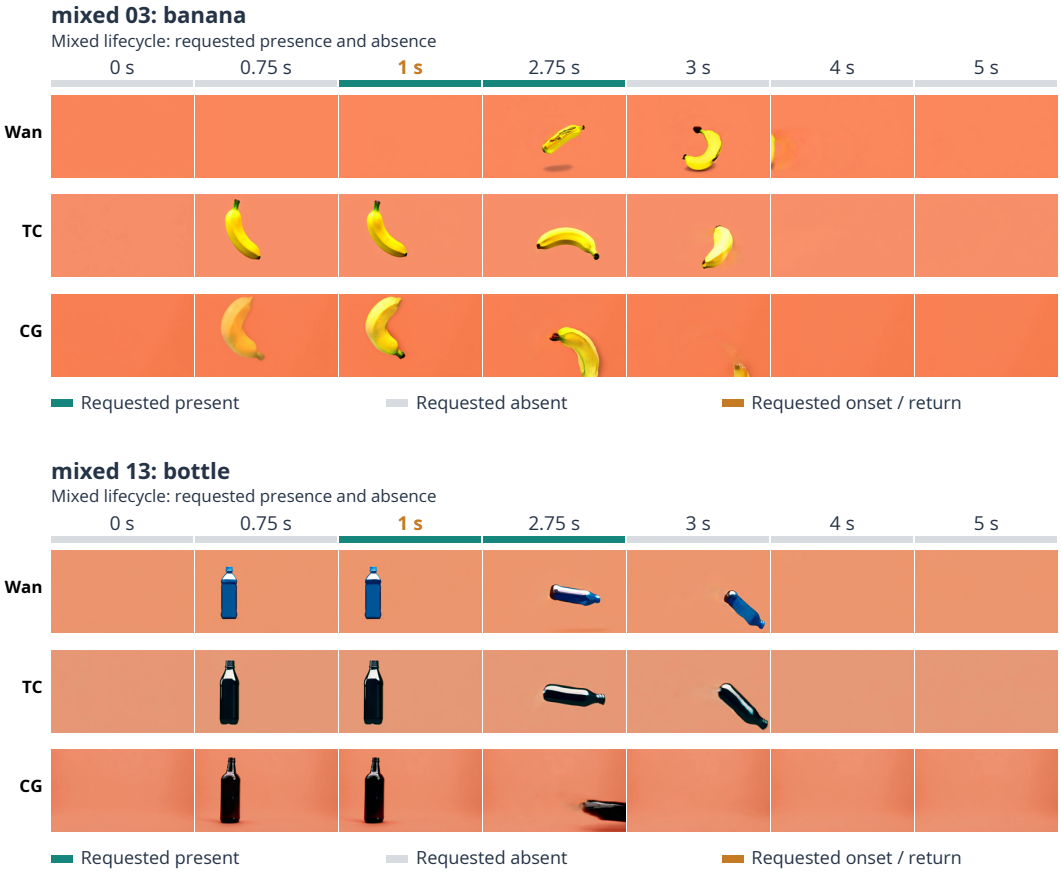



Figure 26: Banana: CG clears the object closer to the requested disappearance time; its appearance remains early. Bottle: CG brings both appearance and disappearance closer to the requested interval; both transitions remain slightly early.

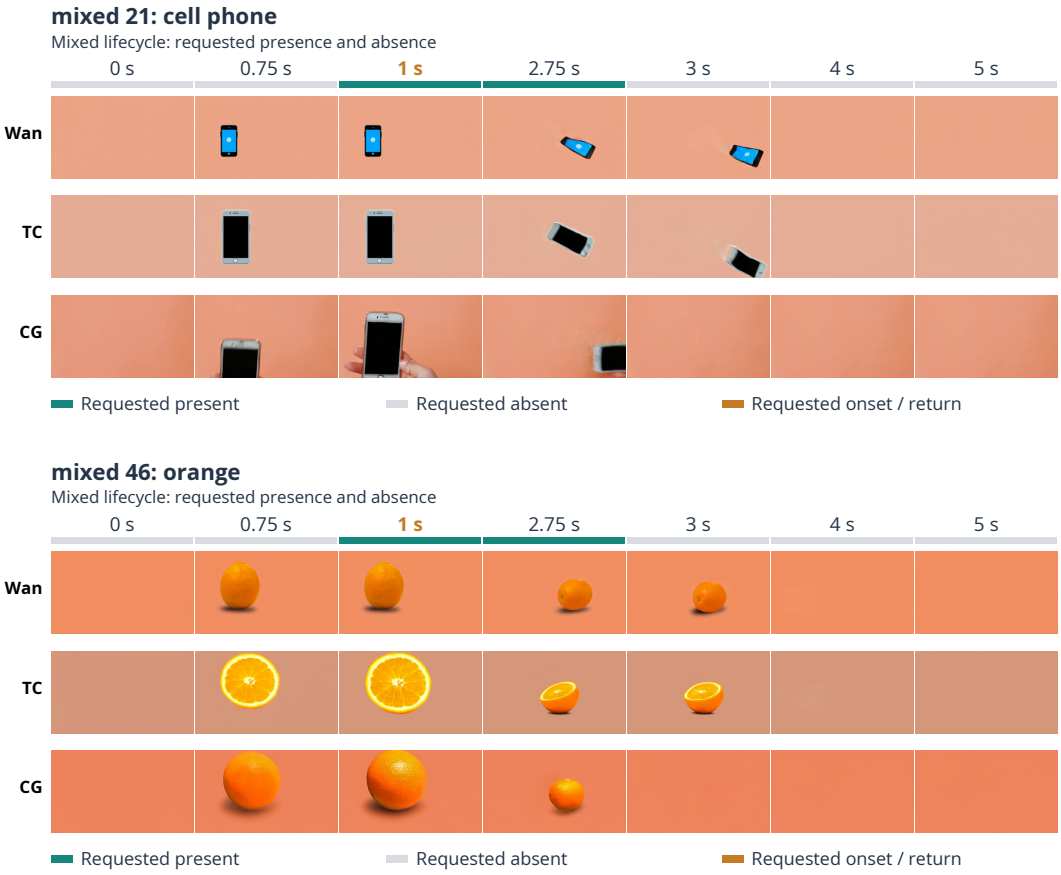



Figure 27: Cell phone: CG brings both appearance and disappearance closer to the requested interval while keeping the scene stable. Orange: CG clears the object closer to the requested disappearance time; all methods enter early.

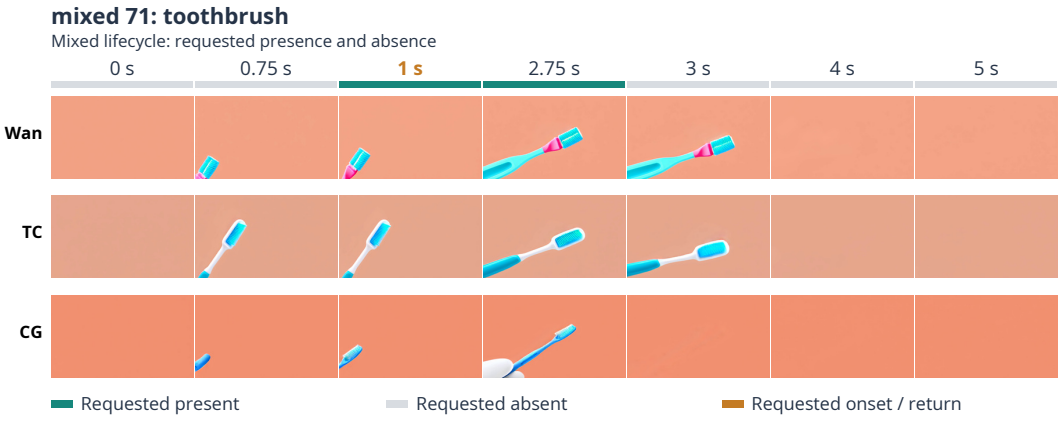



Figure 28: Toothbrush: CG leaves closer to the requested disappearance time; all methods enter early.




Figure 29: **Dog barking.** CG concentrates barking in the requested first second, then returns to a calmer pose. All methods are sampled every 0.25 s; amber marks the requested interval.




Figure 30: **Woman laughing.** CG concentrates pronounced laughter in seconds 1–2 and returns to a relaxed expression afterward. All methods are sampled every 0.25 s; amber marks the requested interval.

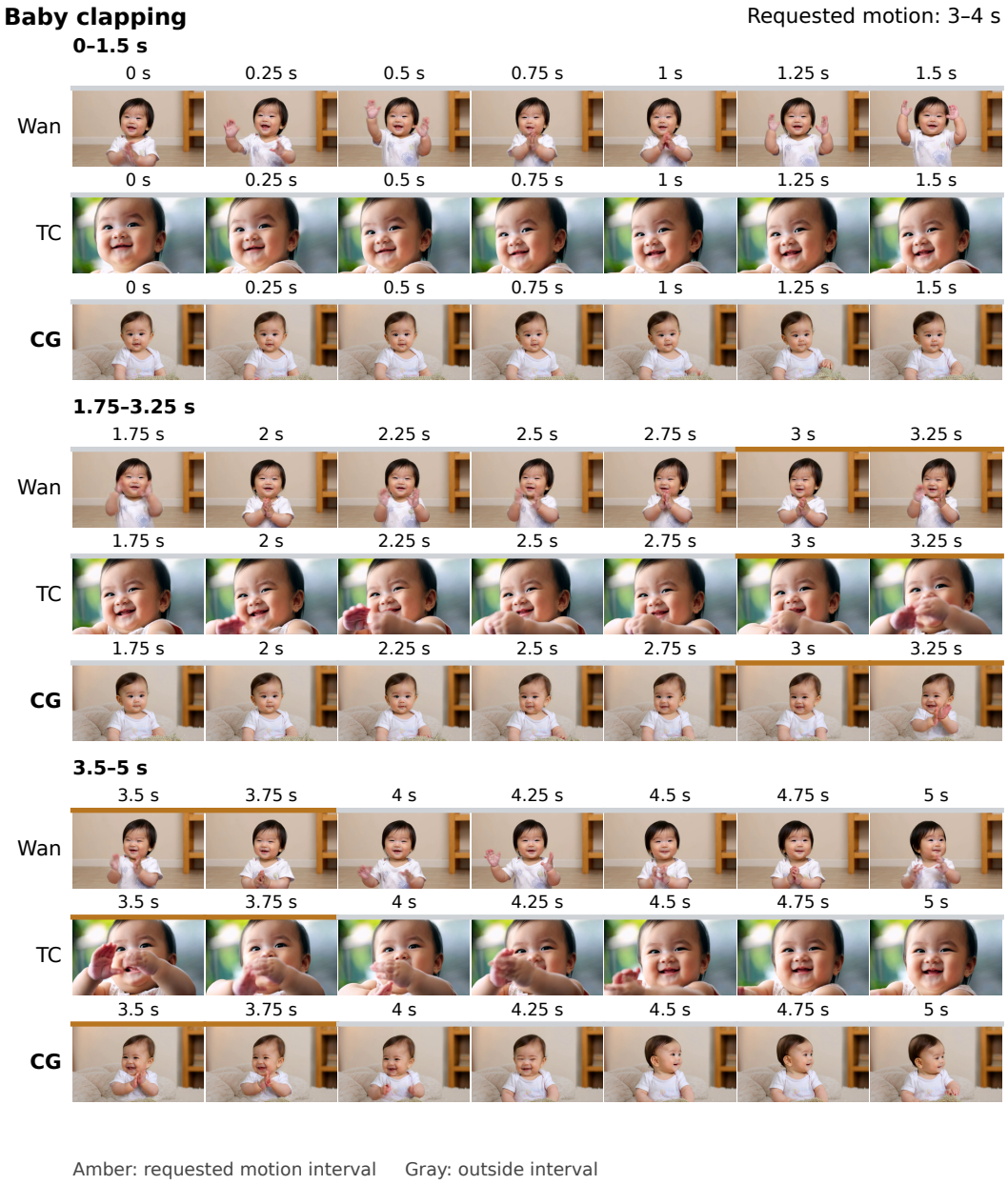



Figure 31: **Baby clapping.** CG delays clapping until the requested 3–4 s interval, with lowered hands beforehand. All methods are sampled every 0.25 s; amber marks the requested interval.

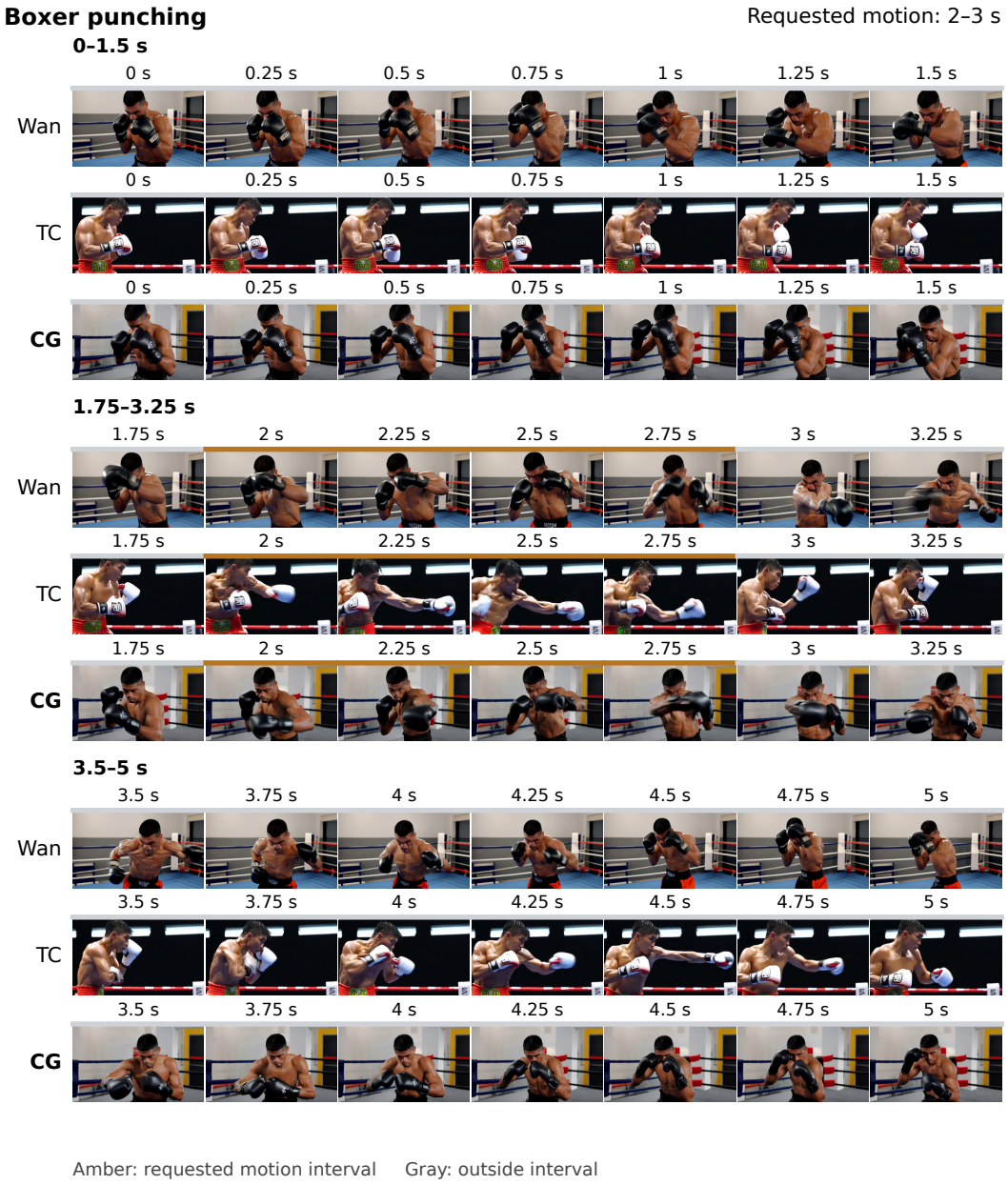



Figure 32: **Boxer punching.** CG concentrates its main punching sequence around the requested 2–3 s interval. All methods are sampled every 0.25 s; amber marks the requested interval.

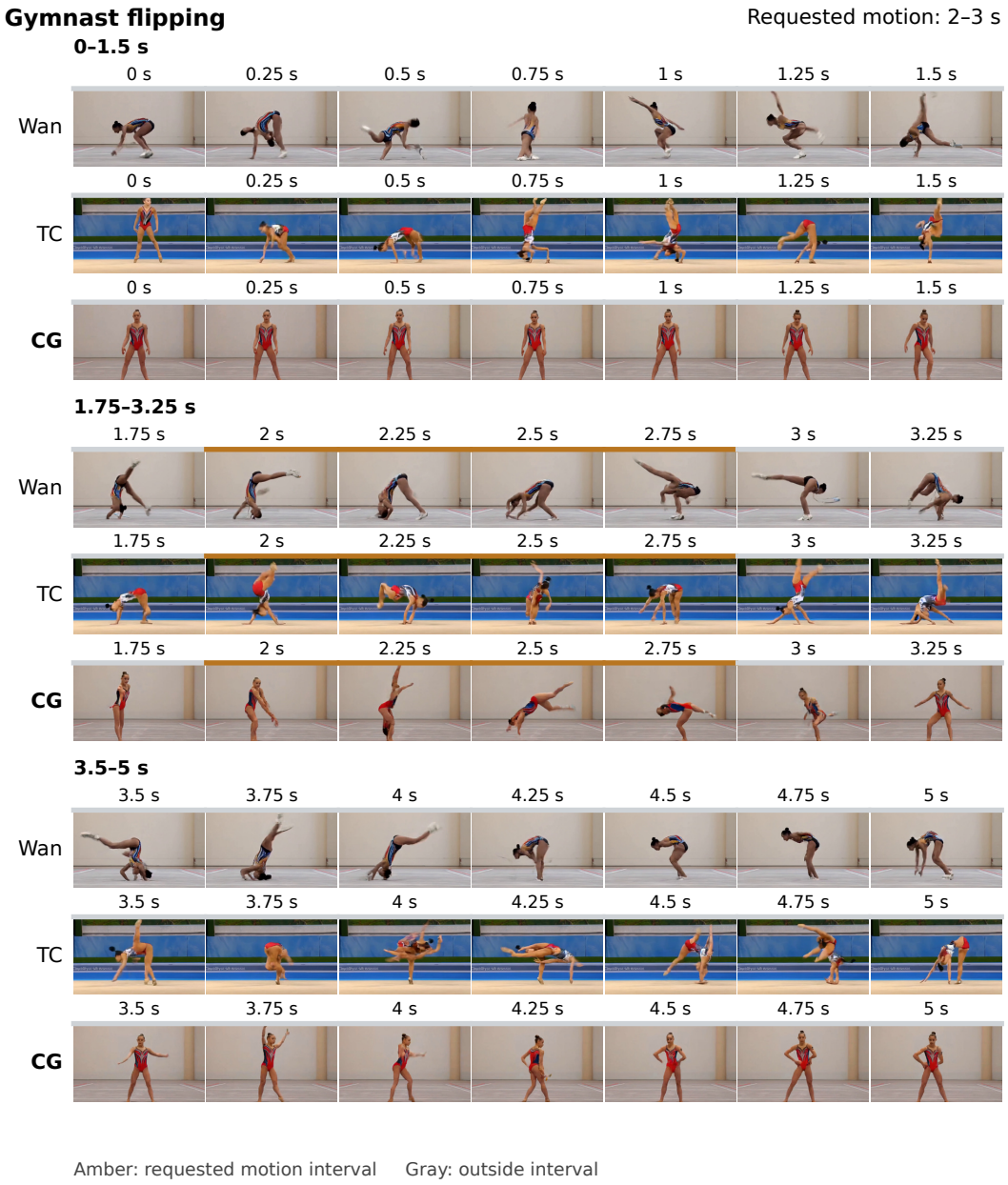



Figure 33: **Gymnast flipping.** CG performs a flip during seconds 2–3, with standing poses before and after. All methods are sampled every 0.25 s; amber marks the requested interval.

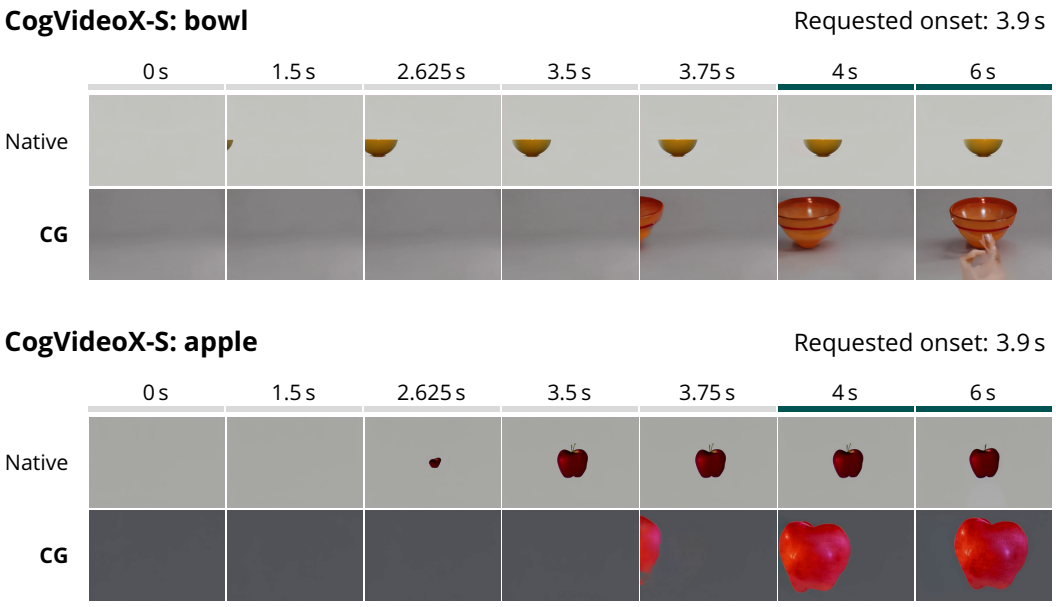



Figure 34: **CogVideoX transfer.** The native bowl begins entering at 1.375 s; the native apple appears at 2.625 s, both before the requested 3.9 s onset. CG delays first visible entry to 3.625 s (bowl) and 3.75 s (apple), within 0.3 s of the request, with recognizable objects at the transition. First visibility is assessed from all 49 original frames, including partial objects at image edges.

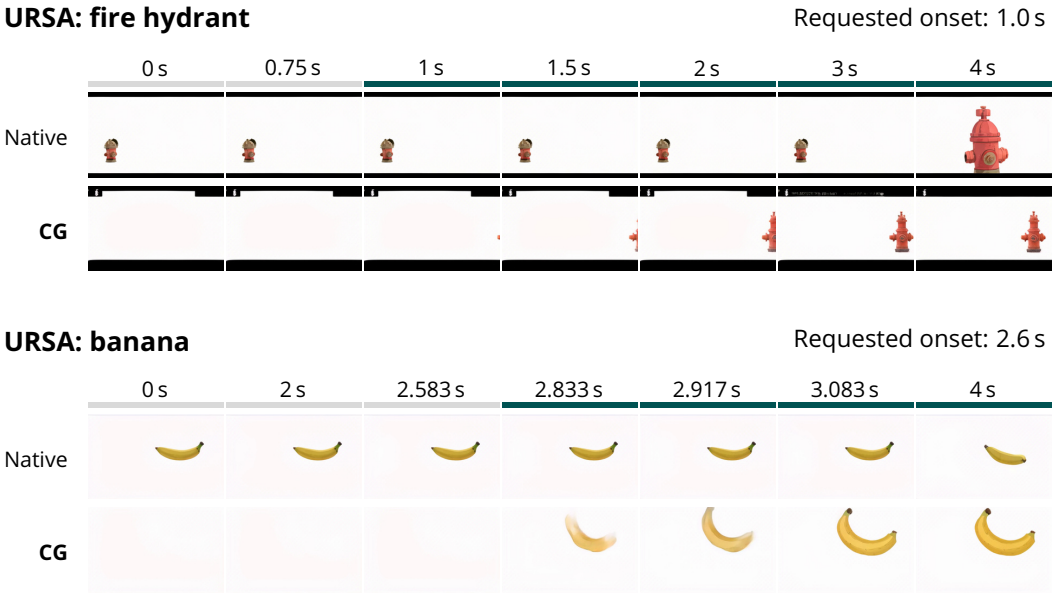



Figure 35: **URSA transfer.** Native generation shows a small hydrant-like object and a banana from the first frame. CG suppresses these early appearances: the hydrant enters from the right around 1 s, while the banana appears around 2.8–2.9 s, near its requested 2.6 s onset.

## N PROMPT-SPAN SELECTION AND BINDING

The branches share the prompt encoding $H_p$ and differ in which text positions they can access. The released registries specify the target and associated descriptions ($J_m^{\mathrm{obj}}$), presence exclusions ($J_m^{\mathrm{exc}}$), and, when used, event-aware absence annotations ($J_m^{\mathrm{evt}}$). Object-only absence masks the concept positions. Event-aware absence additionally masks target-focused and affirmative event phrases, while protecting declared timing, explicit-absence, and continuing-scene phrases from this additional mask.

For example, the mixed airplane prompt masks "appears out of nowhere", "drawing all attention", and "flies away" under event-aware absence, while preserving "An empty scene", the timing phrases, and "is completely out of view" from the event mask. This protection is specific to absence-event masking; presence exclusions are specified separately.

Declared phrases are matched case-insensitively at word boundaries and mapped to token positions using tokenizer offsets. Target and associated conditioning phrases bind at every occurrence; presence exclusions and event annotations require unique matches. The implementation checks native token-ID agreement, rejects truncated or inconsistent bindings, and verifies event annotations against the corresponding prompt.