CUEINSERT: COMPILING RELATIONAL INTENT INTO VERIFIABLE PLANS FOR VIDEO OBJECT INSERTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Video object insertion is inherently a planning-and-generation problem: a system must infer where and how an object should move from high-level intent, and then faithfully realize that plan in the generated video. Existing methods largely focus on visual synthesis or require manually specified low-level controls, leaving intent-driven insertion planning underexplored. We introduce CueInsert, a training-free framework for intent-driven video object insertion via neuro-symbolic planning and plan-conditioned rendering. Given a background video, a reference object, and an insertion intent, CueInsert analyzes scene structure and anchor motion to generate candidate plans over object scale, placement, orientation, and trajectory. Explicit temporal relation predicates then diagnose whether the proposed motion satisfies the intended interaction, while an iterative repair mechanism refines violated trajectories under proposal-preservation and smoothness constraints. The refined candidates are globally reranked and converted into coarse motion and mask conditions for training-free rendering without additional model training or inversion. We further construct DAVIS-Ins, an intent-conditioned benchmark derived from semantically and viewpoint-compatible DAVIS videos for systematic evaluation of planning and rendering, and introduce Insertion Action Coverage (IAC), which measures the fraction of temporal transitions in which the inserted object is present and exhibits residual motion beyond global translation. Experiments demonstrate improved intent consistency, trajectory control, object fidelity, background preservation, and temporal quality over existing video insertion baselines.



Figure 1: **Intent-driven video object insertion with CueInsert.** Different intents produce distinct trajectories and interactions for the same object–background pair while preserving fidelity.

1 INTRODUCTION

Video object insertion aims to integrate a new object into an existing video while preserving visual realism and spatio-temporal consistency Mou et al. (2024); Zhao et al. (2025); Jiang et al. (2025). Compared with image editing Lu et al. (2023); Yang et al. (2024), video insertion additionally requires deciding where the object should appear, how large it should be, how it should move over time, and how its motion should relate to existing actors and scene structure. Although recent diffusion-based video generation and editing methods Chen et al. (2023); Blattmann et al. (2023); Guo et al. (2024); Zhao et al. (2025); Cheng et al. (2024); Ku et al. (2024); Mou et al. (2024) have substantially improved synthesis quality, controllable insertion remains challenging because it is not merely a rendering problem, but also a spatio-temporal planning 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.

Existing approaches typically rely on manually specified layouts, trajectories, or other low-level control signals Mou et al. (2024); Ku et al. (2024); Zhao et al. (2025). Such controls require substantial human effort and do not directly bridge high-level user intent and executable insertion behavior. For example, intents such as *follow*, *approach*, or *parallel* require jointly resolving object scale, placement, orientation, and temporally coherent motion under scene constraints. Even a strong video generator may fail when these upstream controls are implausible or inconsistent. Reliable insertion therefore requires explicitly translating semantic intent into structured, geometrically feasible, and temporally consistent controls before rendering.

This motivates a neuro-symbolic formulation. Multimodal language models OpenAI (2023); Liu et al. (2023); Bai et al. (2023); Chen et al. (2024) provide strong semantic reasoning for interpreting open-ended intents and proposing insertion plans, but cannot reliably guarantee geometric feasibility or temporal consistency. In contrast, symbolic representations can express motion through explicit states such as position, direction, distance, and relative offset, while relation predicates characterize whether an object is truly *following*, *approaching*, or moving *in parallel with* an anchor. These relations can further guide trajectory repair when the initial proposal violates the intended behavior. Thus, neural reasoning provides semantic proposals, while symbolic reasoning represents, diagnoses, and repairs their spatial and temporal structure into executable insertion plans.

Based on this insight, we propose **CueInsert**, an intent-driven framework for controllable video object insertion via neuro-symbolic planning and training-free rendering. Given a background video, reference object, and insertion intent, CueInsert generates candidate plans over scale, placement, orientation, and trajectory, then verifies, repairs, and reranks them using explicit spatiotemporal relations before rendering.

We further construct **DAVIS-Ins**, an intent-conditioned benchmark built from semantically and viewpoint-compatible DAVIS videos, and introduce **Insertion Action Coverage (IAC)** to measure the temporal coverage of inserted-object actions beyond global translation. The selected plan is finally realized through coarse motion and mask guidance without additional training or inversion. As illustrated in Fig. 1, different relational intents lead CueInsert to derive distinct motion plans and interactions for the same object–background pair.

Our contributions are fourfold:

- We propose an intent-driven **neuro-symbolic planning framework** that converts high-level insertion intents into structured spatial-temporal controls.
- We introduce a **relation-aware temporal refinement mechanism** with dynamic anchor tracking, explicit relation predicates, and trajectory repair.
- We develop a **training-free plan-conditioned rendering strategy** using coarse motion and mask guidance without additional training or inversion.
- We construct **DAVIS-Ins** for intent-conditioned video object insertion and introduce **IAC** to quantify the fraction of temporal transitions in which the inserted object is present and exhibits motion beyond global translation.

2 Related Work

Video Editing and Object Insertion. Modern video generation has evolved from latent diffusion and Transformer-based architectures to large-scale video models with improved visual quality and temporal coherence Rombach et al. (2022); Peebles & Xie (2023); Ho et al. (2022); Singer et al. (2023); Blattmann et al. (2023); Kong et al. (2024); Wan Team et al. (2025). Building on these foundations, controllable generation and video editing methods enable flexible manipulation through spatial conditioning, feature injection, and explicit motion control Zhang et al. (2023); Khachatryan et al. (2023); Wang et al. (2024); Wu et al. (2024); Qiu et al. (2026); Ku et al. (2024); Mou et al. (2024); Zhao et al. (2025); Wei et al. (2026). Reference-guided insertion further requires preserving object identity while maintaining spatiotemporal consistency.

Visual Planning and Motion Control. MLLMs enable multimodal reasoning, grounding, planning, and feedback OpenAI (2023); Liu et al. (2023); Surís et al. (2023); Yang et al. (2024), while neuro-symbolic approaches combine neural proposals with explicit structured constraints Mao et al. (2019); Yi et al. (2018). Meanwhile, trajectory- and region-conditioned video generation provides direct

control over object motion Wang et al. (2024); Wu et al. (2024); Zhang et al. (2025); Qiu et al. (2026); Singer et al. (2026). CueInsert connects these directions by translating relational intent into candidate insertion plans, explicitly verifying and repairing their temporal relations, and executing the selected plan through controllable video rendering.

2.1 Intent-Driven Candidate Planning

Candidate planning translates a high-level insertion intent into explicit spatial and motion hypotheses. Scene understanding and anchor observations are shared across candidates, while scale, placement, orientation, and trajectory are instantiated separately to produce diverse candidate plans.

Scene and intent context. An MLLM analyzes the reference object and the background video to extract object attributes, scene semantics, and coarse motion cues. An off-the-shelf detector first localizes the relevant background anchor, which is then tracked throughout the video to obtain $\mathcal{A} = \{B_t^{\mathrm{bg}}\}_{t=1}^{T}$, where $B_t^{\mathrm{bg}} = (\mathbf{a}_t, w_t^{\mathrm{bg}}, h_t^{\mathrm{bg}})$ and $\mathbf{a}_t$ denotes the anchor center. Additional visual queries estimate the facing directions of the reference object and anchor, while the anchor track provides motion and heading cues. Together with the insertion intent p, these observations form a shared planning context $\mathcal{K}$. We consider five behavioral relations: *follow*, *approach*, *avoid*, *parallel*, and *avoid*.

Depth-assisted scale and placement. We estimate a monocular depth map D from the first background frame and define the anchor reference depth as $d_{\mathrm{ref}} = \mathrm{median}_{\mathbf{x}\in B_1^{\mathrm{bg}}} D(\mathbf{x})$. Candidate centers are initialized from regions with depth close to d_{ref}. For candidate n, the MLLM predicts a foreground-to-anchor height ratio $r_h^{(n)}$, giving

$$h^{(n)} = r_h^{(n)} h_1^{\mathrm{bg}}, \qquad w^{(n)} = a^{\mathrm{fg}} h^{(n)}, \tag{1}$$

where a^{fg} is the reference-object aspect ratio. The initial center is then determined jointly by the depth prior, anchor geometry, and intent: *follow* favors a trailing position, *approach* a larger initial distance, and *parallel* a lateral configuration. The resulting scale and placement are further validated by the feedback module.

Intent-conditioned trajectory and orientation. Conditioned on $\mathcal{K}$ and p, the MLLM proposes a center trajectory $\{\tilde{\mathbf{c}}_t^{(n)}\}_{t=1}^{T}$, with the first center anchored to the initialized placement and the object size fixed within each candidate. A binary orientation variable $\tilde{\omega}^{(n)} \in \{0, 1\}$ specifies whether the reference object is horizontally flipped, based on object facing, anchor motion, and the intended relation.

The resulting proposal is $\tilde{\Pi}^{(n)} = (\{\tilde{B}_t^{(n)}\}_{t=1}^{T}, \tilde{\omega}^{(n)})$, where $\tilde{B}_t^{(n)} = (\tilde{\mathbf{c}}_t^{(n)}, w^{(n)}, h^{(n)})$. Repeating the process yields the proposal set $\tilde{\mathcal{P}} = \{\tilde{\Pi}^{(n)}\}_{n=1}^{N}$, which is subsequently verified and repaired by the neuro-symbolic relation module.

2.2 Neuro-Symbolic Feedback

MLLM proposals may still contain implausible scale or violate the intended temporal relation. We therefore combine visual feedback with anchor-relative symbolic verification and trajectory repair before global candidate selection.

Preview-based scale feedback. For each scale hypothesis, we construct a first-frame composite at an intent-conditioned feasible location and use an MLLM judge to assess relative scale plausibility. A bounded search updates the foreground-to-anchor ratio and retains the highest-scoring plausible hypothesis, grounding semantic scale estimation in the target scene.

Anchor-relative relation modeling. Let $\bar{\mathbf{a}}_t$ and $\bar{\mathbf{c}}_t^{(n)}$ denote the normalized anchor and candidate centers, respectively. We characterize their temporal relation by

$$\mathbf{r}_t^{(n)} = \bar{\mathbf{c}}_t^{(n)} - \bar{\mathbf{a}}_t, \qquad d_t^{(n)} = \left\| \mathbf{r}_t^{(n)} \right\|_2, \qquad \Delta \mathbf{r}_t^{(n)} = \mathbf{r}_{t+1}^{(n)} - \mathbf{r}_t^{(n)}. \tag{2}$$

These states support intent-specific predicates Φ_p, where *approach* and *avoid* constrain the evolution of d_t, *follow* and *parallel* additionally enforce motion alignment and relative offset, and *cross* checks

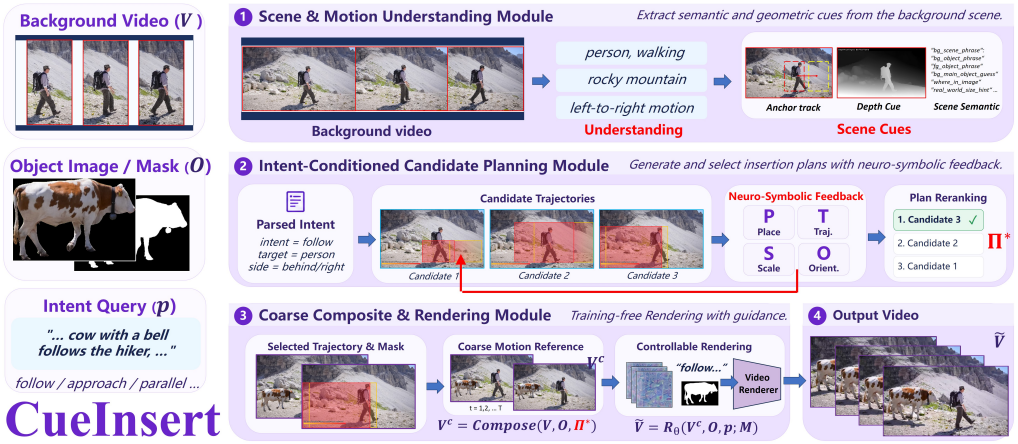

Figure 2: **Overview of CueInsert.** Given a background video $\mathcal{V}$, reference object O, and intent p, CueInsert generates and refines insertion plans through neuro-symbolic reasoning. The selected plan Π^* defines object placement, trajectory, scale, and orientation, and is composed into a coarse video $\mathcal{V}^c = \mathrm{Compose}(\mathcal{V}, O, \Pi^*)$ with trajectory masks $\mathcal{M}$. These conditions guide training-free diffusion rendering to produce the final video $\widetilde{\mathcal{V}}$.

whether the relative trajectory traverses the anchor path. The predicates induce an intent-feasible set

$$\mathcal{F}_p(\mathcal{A}) = \{\Pi \,|\, \Phi_p(\Pi, \mathcal{A}) = \mathrm{pass},\ \Pi \in \Omega_{\mathrm{geo}}\}\,, \tag{3}$$

where Ω_{geo} enforces image-boundary and geometric-validity constraints.

Relation-driven temporal repair. Given a proposal $\tilde{\Pi}^{(n)}$, the relation predicates first identify the violated temporal constraints and the corresponding invalid frame intervals. We then apply a relation-specific repair operator

$$\Pi^{(n)}_{\mathrm{rep}} = \mathcal{R}_p\left(\tilde{\Pi}^{(n)}, \mathcal{A}, \Phi_p\right), \tag{4}$$

which adjusts the candidate center sequence according to the detected violations while preserving the validated scale and initial proposal as much as possible.

For each affected frame, the repaired center is updated as

$$\mathbf{c}_t^{\mathrm{rep}} = \mathrm{Proj}_{\Omega_t}\left(\tilde{\mathbf{c}}_t + \Delta\mathbf{c}_t^{p}\right), \tag{5}$$

where $\Delta\mathbf{c}_t^{p}$ is the relation-specific correction induced by the violated predicate and Ω_t denotes the geometrically valid image region. The resulting trajectory is further smoothed and constrained by bounded positional changes to avoid abrupt motion.

After repair, the same predicates are re-evaluated:

$$\Phi_p\left(\Pi^{(n)}_{\mathrm{rep}}, \mathcal{A}\right) \quad \text{vs.} \quad \Phi_p\left(\tilde{\Pi}^{(n)}, \mathcal{A}\right), \tag{6}$$

and the repaired proposal is retained only when relation consistency is improved.

Relation-aware candidate reranking. Each candidate plan $\tilde{\Pi}^{(n)}$ is first repaired independently. Using the box centers defined by the insertion plan, we optimize the repaired trajectory as

$$\Pi^{(n)} = \underset{\Pi}{\mathrm{argmin}} \sum_{t=1}^{T} \left\|\mathbf{c}_t - \tilde{\mathbf{c}}_t^{(n)}\right\|_2^2 + \lambda_{\mathrm{rel}}\mathcal{L}_{\mathrm{rel}}(\Pi; p) + \lambda_{\mathrm{smooth}} \sum_{t=2}^{T-1} \|\mathbf{c}_{t+1} - 2\mathbf{c}_t + \mathbf{c}_{t-1}\|_2^2\,, \tag{7}$$

subject to image-boundary and per-axis displacement constraints. Here, $\tilde{\mathbf{c}}_t^{(n)}$ and $\mathbf{c}_t$ denote the original and repaired box centers at frame t, respectively, while $\mathcal{L}_{\mathrm{rel}}(\Pi; p)$ measures the soft violation of the requested relation p. The first term preserves the original proposal, and the last term penalizes temporal acceleration. A repaired trajectory is retained only when both the overall objective and the

Algorithm 1 Neuro-Symbolic Insertion Planning and Selection

Require: Background video $\mathcal{V}$, reference object O, intent p, candidate budget N, scale-trial budget M_s
Ensure: Selected plan Π^*
Build planning context $\mathcal{K}$ and estimate D, d_{ref}, and anchor trajectory $\{\bar{\mathbf{a}}_t\}_{t=1}^{T}$.
$\mathcal{P} \leftarrow \emptyset$.
for $n = 1, \ldots, N$ **do**
  Initialize candidate placement and scale using depth and MLLM cues.
  Refine the scale through at most M_s preview-feedback trials.
  Generate trajectory $\{\bar{\mathbf{c}}_t^{(n)}\}_{t=1}^{T}$ and coarse orientation.
  Compute anchor-relative states and evaluate $\Phi_p(\tilde{\Pi}^{(n)}, \mathcal{A})$.
  if relation violations are detected **then**
    Apply relation-specific repair $\Pi_{\text{rep}}^{(n)} \leftarrow \mathcal{R}_p(\tilde{\Pi}^{(n)}, \mathcal{A}, \Phi_p)$.
    Re-evaluate the relation predicates.
  else
    $\Pi_{\text{rep}}^{(n)} \leftarrow \tilde{\Pi}^{(n)}$.
  end if
  Construct refined plan $\Pi^{(n)}$ and compute reranking score $S^{(n)}$.
  $\mathcal{P} \leftarrow \mathcal{P} \cup \{\Pi^{(n)}\}$.
end for
$\Pi^* \leftarrow \arg\max_{\Pi^{(n)} \in \mathcal{P}} S^{(n)}$.
return Π^*.

relation violation decrease. The corresponding relation predicate is then re-evaluated, since a lower soft violation alone does not guarantee relation satisfaction.

Global selection operates on the repaired plans and their composited previews. A candidate $\Pi^{(n)}$ is considered admissible, denoted by $\Pi^{(n)} \in \mathcal{A}$, if all boxes remain inside the image and a vision-language judge assigns a semantic score $S_{\text{sem}}^{(n)} \in \{1, \ldots, 10\}$ above the rejection threshold. Among admissible candidates, we select

$$\Pi^* = \underset{\Pi^{(n)} \in \mathcal{A}}{\text{lexmin}} \left(\mathbf{1}\left[\neg \text{cert}_p\left(\Pi^{(n)}\right)\right], -S_{\text{sem}}^{(n)}, \mathcal{L}_{\text{rel}}\left(\Pi^{(n)}; p\right), \frac{1}{T}\sum_{t=1}^{T} \left\| \mathbf{c}_t^{(n)} - \tilde{\mathbf{c}}_t^{(n)} \right\|_2 \right), \tag{8}$$

where lexmin compares entries from left to right. Here, $\text{cert}_p(\Pi^{(n)})$ indicates whether the repaired plan satisfies the relation predicate for intent p, $S_{\text{sem}}^{(n)}$ is the semantic score of the repaired preview, $\mathcal{L}_{\text{rel}}(\Pi^{(n)}; p)$ is the residual relation violation, and the last term measures the mean deviation from the original trajectory. Thus, the ranking prioritizes certified plans, followed by higher semantic scores, lower relation violation, and smaller repair magnitude. If no certified candidate is admissible, the best semantically acceptable uncertified plan is retained for rendering.

2.3 Plan-Conditioned Controllable Rendering

The selected plan Π^* is converted into a coarse video $\mathcal{V}^c = \text{Compose}(\mathcal{V}, O, \Pi^*)$ and trajectory masks $\mathcal{M}$. We use HunyuanVideo as the pretrained video diffusion backbone, with the conditioning interface adapted from UniVideo Wei et al. (2026), and adopt the sampling principle of Time-to-Move Singer et al. (2026). The native text, object, and video conditions remain unchanged, while $\mathcal{V}^c$ provides an additional training-free motion prior during sampling.

Let $z^c = \mathcal{E}(\mathcal{V}^c)$ be the VAE latent and m the corresponding latent-space mask. Using shared Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$, we construct

$$z_{\text{ref}}(\sigma) = (1 - \sigma)z^c + \sigma\epsilon, \qquad z_{k_s} = z_{\text{ref}}(\sigma_{k_s}), \tag{9}$$

which initializes sampling without inversion. During early denoising, the planned object region is anchored to this coarse reference and subsequently released:

$$z_{k+1} = \begin{cases} (1 - m) \odot \widetilde{z}_{k+1} + m \odot z_{\text{ref}}(\sigma_{k+1}), & k_s \le k < k_r, \\ \widetilde{z}_{k+1}, & k_r \le k < K, \end{cases} \tag{10}$$

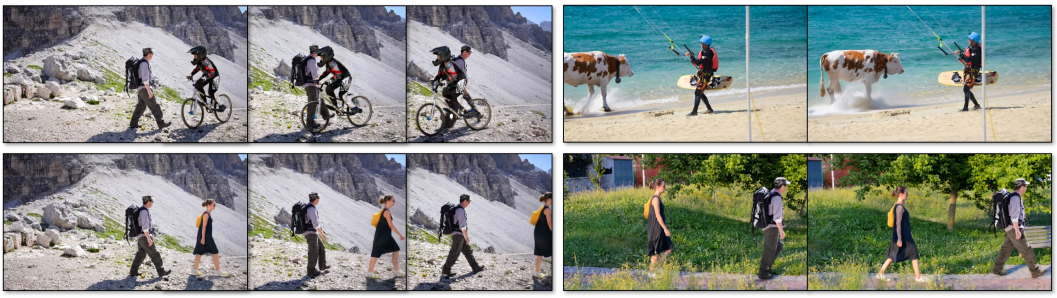

Figure 3: Representative examples from DAVIS-Ins across different background–object pairs.



Figure 4: Residual-motion illustration for IAC. Blue contours show the observed object, yellow dashed contours indicate translation-only predictions, and red arrows show residual displacements.

Table 1: Evaluation metrics used in DAVIS-Ins.

Category	Metric(s)
Object Fidelity	CLIP-I Radford et al. (2021), DINO-I Caron et al. (2021)
Text Alignment	CLIP-T Radford et al. (2021)
Visual Quality	Aesthetic Quality Huang et al. (2024)
Temporal Consistency	Subject Consistency, Background Consistency Huang et al. (2024)
Insertion Success	IAC (Ours)

where $\widetilde{z}_{k+1}$ is the native model update and k_r denotes the release step. After release, HunyuanVideo freely refines appearance and motion under its native conditioning, yielding the final video $\widetilde{\mathcal{V}} = \mathcal{D}(z_K)$, where $\mathcal{D}$ denotes the pretrained VAE decoder.

3 DAVIS-Ins Benchmark

Construction We construct **DAVIS-Ins** from DAVIS-2016 Perazzi et al. (2016) by selecting nine side-view videos with compatible scene semantics and medium-scale foreground subjects, each sampled to 33 frames. For every selected video used as the background, each of the remaining eight videos provides a segmented foreground object as the reference, yielding $9 \times 8 = 72$ ordered background–object pairs. Each pair is instantiated with four relational intents—*follow*, *avoid*, *parallel*, and *approach*—and two retained plans per intent, resulting in $72 \times 4 \times 2 = 576$ intent-conditioned insertion instances. Representative examples are shown in Fig. 3, with further construction details provided in Appendix A.

Evaluation Protocol We evaluate generated videos using the metrics summarized in Table 1. CLIP-I and DINO-I measure reference-object fidelity, while CLIP-T evaluates semantic alignment with the intent-conditioned prompt. For video-level quality and temporal behavior, we adopt VBench Huang et al. (2024), including Aesthetic Quality, Subject Consistency, and Background Consistency. These metrics jointly characterize perceptual quality, cross-frame consistency, and the overall amount of visual motion.

While Dynamic Degree captures overall video dynamics, it cannot distinguish rigid object translation from internal object dynamics. Inspired by the residual-motion analysis in GeCo Gu et al. (2026), we introduce **Insertion Action Coverage (IAC)** to quantify the temporal coverage of inserted-object actions beyond global translation. Specifically, IAC measures the fraction of adjacent frame pairs in which the inserted object is present and exhibits residual motion above a fixed threshold. Fig. 4 illustrates how subtracting the prescribed translation reveals local motion of the inserted object.

For consecutive frames, let $F_t(\mathbf{x}) \in \mathbb{R}^2$ denote the optical flow at pixel $\mathbf{x}$ and $\mathbf{c}_t$ the prescribed object center, both expressed in pixel units at the flow resolution. We subtract the uniform translation $\Delta\mathbf{c}_t = \mathbf{c}_{t+1} - \mathbf{c}_t$ at each pixel and normalize the mean residual magnitude by the object bounding-box

Table 2: Quantitative comparison with state-of-the-art video object insertion and editing methods. The best result is shown in **bold**, and the second-best is underlined.

Metric	ICVE	ReVideo	VideoCoF	UniVideo	AnyV2V	D-Insert	Senorita	CueInsert
CLIP-I ↑	0.8213	0.9077	0.8091	0.8528	0.8872	0.8679	0.8725	**0.9159**
CLIP-T ↑	0.3063	0.3208	0.3088	0.3201	0.3243	**0.3369**	0.2956	0.3323
DINO-I ↑	0.5925	0.6824	0.6127	0.6072	0.6740	0.6205	0.6647	**0.7039**
Aesthetic Quality ↑	0.5474	0.5310	**0.5560**	0.5505	0.5212	0.5074	0.5250	0.5506
BG Consistency ↑	0.9433	0.9407	0.9159	**0.9553**	0.9129	0.9268	0.9420	0.9413
Subject Consistency ↑	0.9402	0.9177	0.9144	0.9485	0.8709	0.9087	0.9164	**0.9515**
IAC ↑	0.5764	0.6771	0.5278	0.7361	0.6853	0.7604	0.7031	**0.8160**
Avg. ↑	0.6753	0.7111	0.6635	0.7101	0.6965	0.7041	0.7028	**0.7445**

diagonal:

$$r_t = \frac{1}{|M_t|\sqrt{w_t^2 + h_t^2}} \sum_{\mathbf{x} \in M_t} \left\| \underbrace{F_t(\mathbf{x}) - \Delta \mathbf{c}_t}_{F_t^{\text{res}}(\mathbf{x})} \right\|_2, \tag{11}$$

where M_t is the object region and (w_t, h_t) its bounding-box size.

Rather than directly averaging the residual magnitude, we use it to determine whether residual object motion occurs at each temporal transition:

$$a_t = \begin{cases} 1, & r_t > \tau, \\ 0, & r_t \le \tau, \end{cases} \qquad \text{IAC} = \frac{1}{T-1} \sum_{t=1}^{T-1} a_t \in [0, 1]. \tag{12}$$

Here, τ suppresses small residuals caused by optical-flow noise or imperfect translation compensation. If the inserted object is missing or cannot be reliably localized in either frame of a pair, we set $a_t = 0$. Higher IAC indicates greater temporal coverage of object actions, without rewarding larger residual-motion magnitudes. Further analysis is provided in Appendix C.2.

4 Experiments

Settings. We evaluate CueInsert on DAVIS-Ins and compare it with ICVE Liao et al. (2025), ReVideo Mou et al. (2024), VideoCoF Yang et al. (2026), UniVideo Wei et al. (2026), AnyV2V Ku et al. (2024), DreamInsert Zhao et al. (2025), and Senorita Zi et al. (2025) using their official implementations and recommended settings when available. For rendering, CueInsert adopts a HunyuanVideo-based diffusion backbone following the UniVideo Wei et al. (2026) formulation, without additional training or fine-tuning. All experiments are conducted on an NVIDIA RTX PRO 6000 Blackwell Server Edition GPU.

4.1 Quantitative Comparison

We compare CueInsert with representative video object insertion and controllable editing methods on DAVIS-Ins, as reported in Table 2. CueInsert achieves the best average score of 0.7445, outperforming the second-best ReVideo (0.7111) by 4.7%. It also ranks first in CLIP-I (0.9159 vs. 0.9077), DINO-I (0.7039 vs. 0.6824), Subject Consistency (0.9515 vs. 0.9485), and IAC (0.8160 vs. 0.7604). Notably, CueInsert improves IAC by 7.3% over the second-best method, demonstrating substantially stronger intent-aware interaction while maintaining competitive CLIP-T, consistency, and aesthetic quality.

4.2 Qualitative comparison.

As shown in Fig. 5, CueInsert better preserves object identity, generates plausible actions, and follows the specified interaction intent. The Lucia→Kite-walk case highlights intent-driven spatial interaction, while Bear→Rhino examines long- term object and action preservation. Additional qualitative results are provided in Appendix D.

Comparison with SOTA. In Lucia→Kite-walk, frames 1, 15, and 30 are shown. CueInsert realizes the *approach* intent with stable insertion and coherent motion; ReVideo becomes increasingly blurry, while UniVideo largely keeps the object stationary. In Bear→Rhino, where the first and final frames

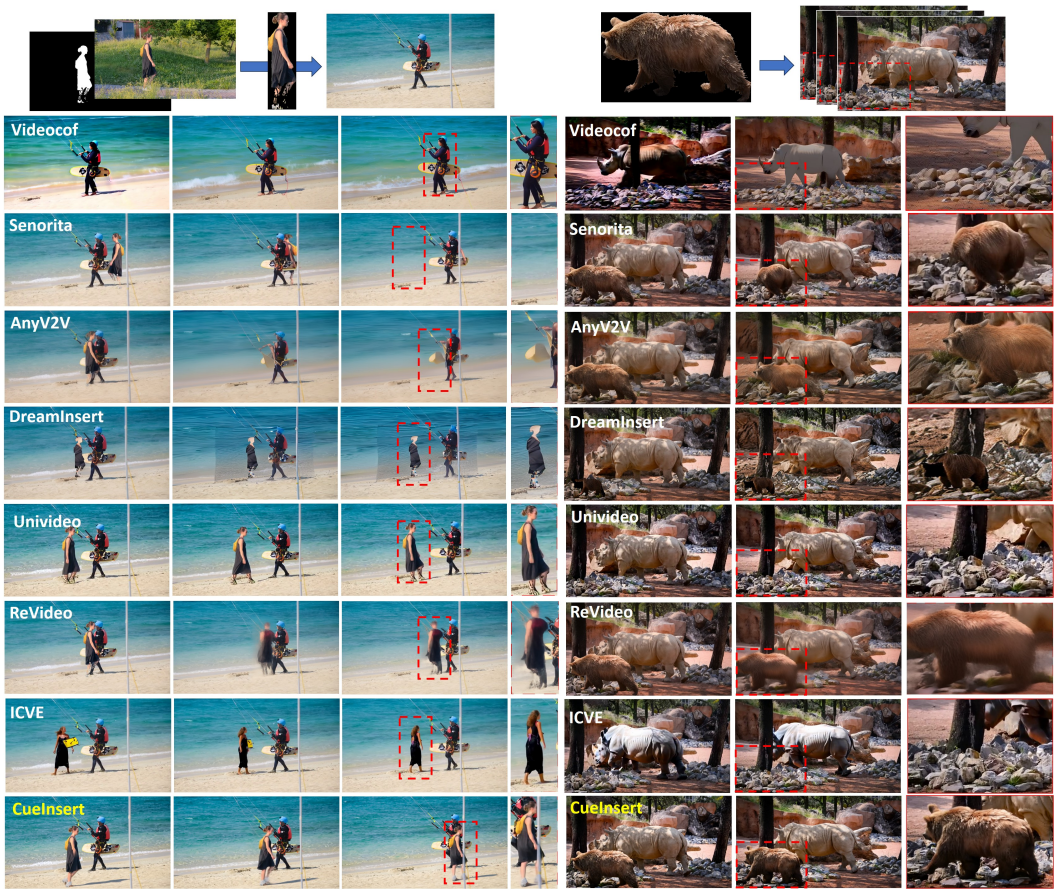

Figure 5: **Qualitative comparison on DAVIS-Ins.** Results are shown for Lucia→Kite-walk under the *approach* intent and Bear→Rhino. CueInsert preserves object fidelity, produces plausible object actions, and better aligns the resulting interaction with the specified intent.

are shown, CueInsert further maintains both object fidelity and action throughout the video, enabling more coherent long-term interaction with the scene.

4.3 Planning Evaluation

Since CueInsert explicitly separates insertion planning from video rendering, we evaluate the generated plans before synthesis. For each background–object–intent case, the planner produces five candidate plans, which are checked by the corresponding temporal relation predicates after symbolic refinement when applicable. We report the *Relation Satisfaction Rate* (RSR), i.e., the percentage of evaluable candidates that satisfy the intended relation. We further measure *Feasible-1* and *Feasible-2*, denoting the percentages of cases containing at least one and two predicate-certified plans, respectively, as well as the average number of valid plans per case.

As shown in Table 3, 83.97% of the evaluated cases contain at least one predicate-certified candidate and 68.22% contain at least two, with an average of 2.38 valid plans per case, supporting our set-valued planning formulation. Among the evaluated relations, *follow* achieves the highest candidate-level satisfaction rate of 59.70%, whereas *parallel* remains more challenging due to its stricter relative-motion constraints. Overall, these results show that CueInsert frequently produces multiple relation-consistent plans before synthesis, providing feasible candidates for subsequent rendering.

4.4 User Study

To complement automatic metrics, we conduct a perceptual study with 20 participants comparing CueInsert against ICVE, ReVideo, Senorita, and UniVideo. For each case, participants rate videos

Table 3: Evaluation of insertion-plan quality before video rendering.

Intent	RSR ↑	Feas. 1 ↑	Feas. 2 ↑	Valid/ Case ↑
Approach	48.53	98.53	77.94	2.43
Avoid	55.59	86.76	72.06	2.78
Follow	59.70	80.60	76.12	2.99
Parallel	33.43	65.22	47.83	1.64
Overall	**49.30**	**82.72**	**68.38**	**2.45**

Table 4: User study results averaged over the evaluated cases. Higher is better.

Method	Object Fidelity	Motion Natural.	Intent Consist.	Inter. Plausib.	Overall Quality
CueInsert	**4.62**	**4.36**	**4.41**	**4.39**	**4.41**
ICVE	2.67	2.14	1.65	1.53	1.95
ReVideo	3.02	2.51	2.10	1.91	2.31
Senorita	3.47	2.42	2.09	2.09	2.50
UniVideo	4.19	3.61	3.11	3.11	3.52

Intent	RSR Before ↑	RSR After ↑	ΔRSR ↑	Repair Success ↑	D_{mean} ↓
Approach	39.71	48.53	+8.82	14.63	42.47
Avoid	43.82	55.59	+11.76	20.94	36.30
Follow	5.99	59.70	+53.71	46.22	46.62
Parallel	1.62	33.43	+31.81	26.32	75.48
Overall	**24.28**	**49.30**	**+25.01**	**27.97**	**52.63**



Figure 6: Ablation of symbolic relation repair. The table reports relation satisfaction before and after repair, while the qualitative example shows a representative case.

generated from the same background–object pair and insertion intent on a five-point Likert scale in terms of *object fidelity*, *motion naturalness*, *intent consistency*, *interaction plausibility*, and *overall quality*. Method identities are hidden and the presentation order is randomized. Table 4 summarizes the results, where CueInsert achieves the highest scores across all criteria, particularly in intent consistency and interaction plausibility.

4.5 Ablation Study

We conduct ablation studies to analyze the contributions of both the planning and rendering components of CueInsert.

Ablation on Relation-Aware Planning. We evaluate symbolic relation repair by comparing the relation satisfaction rate (RSR) before and after refinement, together with the repair success rate and trajectory displacement D_{mean}. As shown in Table 6, repair improves the overall RSR from 24.28% to 49.30%, with 27.97% of initially failed plans converted to predicate-satisfying ones. The gains are particularly large for *follow* and *parallel*, demonstrating the benefit of explicit relation-aware refinement. Further metric definitions and per-intent analysis are provided in Appendix C.1.

Plan-conditioned rendering. We compare CueInsert with the native UniVideo renderer under the same HunyuanVideo backbone. As shown in Fig. 6, UniVideo produces plausible local object actions but may fail to realize the intended spatial interaction, whereas CueInsert follows the planned trajectory while preserving natural object motion. This demonstrates the benefit of explicit plan conditioning for coordinated motion control.

5 Conclusion

We presented **CueInsert**, an intent-driven framework that converts high-level insertion intent into executable spatial-motion plans through neuro-symbolic planning, followed by training-free video rendering. We further constructed **DAVIS-Ins** for systematic evaluation under diverse interaction intents. Experiments show that explicit planning improves intent consistency and motion control while preserving visual fidelity and temporal coherence. These results support planning-before-generation as an effective paradigm for controllable video object insertion.

Ethics Statement

CueInsert enables controllable video manipulation, which may be used for creative applications but could also facilitate misleading synthetic content. We encourage responsible use and appropriate disclosure of generated media.

Reproducibility Statement

Implementation details, dataset construction, evaluation protocols, and additional results are provided in the appendix.

AI Assistance

Generative AI tools were used solely to assist with language polishing and improving the clarity of the manuscript. They were not used to generate research ideas, experimental results, analyses, or conclusions. All technical content and final manuscript decisions were reviewed and approved by the authors.

References

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. URL https://arxiv.org/abs/2308.12966.

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, 2023.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, 2021.

Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, Chao Weng, and Ying Shan. VideoCrafter1: Open diffusion models for high-quality video generation, 2023.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Jiaxin Cheng, Tianjun Xiao, and Tong He. Consistent video-to-video transfer using synthetic dataset. In *International Conference on Learning Representations*, 2024.

Leslie Gu, Junhwa Hur, Charles Herrmann, Fangneng Zhan, Todd Zickler, Deqing Sun, and Hanspeter Pfister. Geco: Evaluating geometric consistency for video generation via motion and structure. In *European Conference on Computer Vision*, 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 *International Conference on Learning Representations*, 2024.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey A. Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen video: High definition video generation with diffusion models, 2022.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. 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.

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

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.

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, 2024. URL https://arxiv.org/abs/2412.03603.

Max Ku, Cong Wei, Weiming Ren, Huan Yang, and Wenhu Chen. AnyV2V: A tuning-free framework for any video-to-video editing tasks. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Xinyao Liao, Xianfang Zeng, Ziye Song, Zhoujie Fu, Gang Yu, and Guosheng Lin. In-context learning with unpaired clips for instruction-based video editing, 2025. URL https://arxiv.org/abs/2510.14648.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Shilin Lu, Yanzhu Liu, and Adams Wai-Kin Kong. TF-ICON: Diffusion-based training-free cross-domain image composition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2294–2305, 2023.

Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, and Jiajun Wu. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. In *International Conference on Learning Representations*, 2019.

Chong Mou, Mingdeng Cao, Xintao Wang, Zhaoyang Zhang, Ying Shan, and Jian Zhang. ReVideo: Remake a video with motion and content control. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

OpenAI. GPT-4V(ision) system card. OpenAI system card, 2023. URL https://cdn.openai.com/papers/GPTV_System_Card.pdf.

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

Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus H. Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 724–732, 2016.

Haonan Qiu, Zhaoxi Chen, Zhouxia Wang, Yingqing He, Menghan Xia, and Ziwei Liu. FreeTraj: Tuning-free trajectory control via noise guided video diffusion. *International Journal of Computer Vision*, 134(4):165, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021.

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*, pp. 10684–10695, 2022.

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 *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=0xO9OSYVw5.

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 *International Conference on Learning Representations*, 2023.

Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Wan Team, 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.

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*, pp. 114:1–114:11. ACM, 2024.

Cong Wei, Quande Liu, Zixuan Ye, Qiulin Wang, Xintao Wang, Pengfei Wan, Kun Gai, and Wenhu Chen. Univideo: Unified understanding, generation, and editing for videos. In *International Conference on Learning Representations*, 2026.

Weijia Wu, Zhuang Li, Yuchao Gu, Rui Zhao, Yefei He, David Junhao Zhang, Mike Zheng Shou, Yan Li, Tingting Gao, and Di Zhang. DragAnything: Motion control for anything using entity representation. In *European Conference on Computer Vision*, 2024. URL https://arxiv.org/abs/2403.07420.

Ling Yang, Zhaochen Yu, Chenlin Meng, Minkai Xu, Stefano Ermon, and Bin Cui. Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Xiangpeng Yang, Ji Xie, Yiyuan Yang, Yue Ma, Yan Huang, Min Xu, and Qiang Wu. Videocof: Unified video editing with temporal reasoner. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Joshua B. Tenenbaum. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023.

Zhongwei Zhang, Fuchen Long, Zhaofan Qiu, Yingwei Pan, Wu Liu, Ting Yao, and Tao Mei. MotionPro: A precise motion controller for image-to-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://arxiv.org/abs/2505.20287.

Qi Zhao, Zhan Ma, and Pan Zhou. Dreaminsert: Zero-shot image-to-video object insertion from a single image, 2025.

Bojia Zi, Penghui Ruan, Marco Chen, Xianbiao Qi, Shaozhe Hao, Shihao Zhao, Youze Huang, Bin Liang, Rong Xiao, and Kam-Fai Wong. Señorita-2M: A high-quality instruction-based dataset for general video editing by video specialists, 2025. URL https://arxiv.org/abs/2502.06734.

A Additional Details of DAVIS-Ins

DAVIS-Ins is constructed from nine DAVIS-2016 videos: *bear*, *bmx-bumps*, *camel*, *cows*, *hike*, *kite-walk*, *lucia*, *rhino*, and *rollerblade*. These videos cover diverse foreground subjects, including humans, animals, and bicycle-related motion, as well as varied outdoor scenes such as roads, grasslands, and natural environments. They also exhibit heterogeneous motion patterns, ranging from walking and running to riding, group movement, and articulated animal motion.

We select these videos to provide semantically plausible yet sufficiently diverse cross-video insertion pairs. In particular, the selected subjects are predominantly visible from side or oblique views, occupy moderate spatial extents, and exhibit clear foreground motion, making them suitable for evaluating relational behaviors such as *follow*, *avoid*, *parallel*, and *approach*.

B Implementation Details

Planning and relation repair. For each input, CueInsert generates five candidate plans over 33 frames. Dynamic anchor tracking is performed at 24 fps. Relation repair uses position-based optimization with $\lambda_{\text{rel}} = 256$ and $\lambda_{\text{smooth}} = 4$. The maximum per-axis trajectory modification is limited to 0.15 of the image width or height. We use five internal control points, an initial search step of 0.22, two refinement rounds, and at most 200 objective evaluations per candidate.

Relation-aware reranking. We use Qwen3-VL-30B-Instruct for semantic assessment with temperature 0.1 and a maximum output length of 256 tokens. Six uniformly sampled frames are used for each visual assessment. A semantic score no greater than 3 is treated as unacceptable. All candidates are evaluated and ranked by relation certification, semantic score, residual relation violation, and repair magnitude.

Controllable rendering. We use UniVideo with the HunyuanVideo backbone as the rendering backend. Videos are rendered at 480×832 resolution using 30 denoising steps, with guidance scale 7.0, image guidance scale 2.0, timestep shift 7.0, and seed 42. The selected plan is converted into a coarse composite video and moving object masks, which guide the adapted TTM sampling procedure without additional training or inversion.

C Additional Analysis

C.1 Relation-Aware Planning

We provide additional details on the relation-repair ablation in Table 6. For each candidate plan, the corresponding temporal predicate is evaluated before and after symbolic refinement. The relation satisfaction rate (RSR) is defined over evaluable candidates, with plans of *unknown* predicate status excluded from the denominator. *Repair Success* measures the fraction of originally failed plans that become predicate-satisfying after repair. We additionally report D_{mean}, the mean center displacement in pixels between the original and repaired trajectories for accepted modifications.

Across the four repair-enabled intents, RSR increases from 24.28% to 49.30%, corresponding to an improvement of 25.01 percentage points. Among originally failed candidates, 27.97% are successfully converted into predicate-satisfying plans. The largest improvements occur for *follow* and *parallel*, whose RSR increases by 53.71 and 31.81 percentage points, respectively. These relations depend strongly on relative motion with respect to the dynamic anchor, and therefore benefit most from explicit temporal constraints. For *approach* and *avoid*, RSR improves by 8.82 and 11.76 points, respectively.



Figure 7: **IAC captures insertion-specific action beyond generic motion smoothness.** Although all methods obtain similarly high VBench Motion Smoothness scores, IAC more clearly distinguishes sustained object-level action from failed or unstable insertions. The two values shown at the lower left of each example denote Motion Smoothness (white) and IAC (yellow), respectively.

The average displacement introduced by accepted repairs is 52.63 pixels overall. This metric quantifies how much the original semantic proposal is modified to improve relation satisfaction, rather than treating successful repair as unconstrained trajectory regeneration.

C.2 Discussion of Motion-Related Metrics

IAC vs. generic motion metrics. VBench Huang et al. (2024) motion metrics provide useful complementary measurements but do not directly assess whether the inserted object exhibits meaningful object-level action. For example, *Dynamic Degree* can be affected by camera or background motion, while *Motion Smoothness* may remain high even when the inserted object is absent, fades over time, or is only rigidly translated along a prescribed trajectory.

As illustrated in Fig. 7, CueInsert, Senorita Zi et al. (2025), AnyV2V Ku et al. (2024), and ICVE Liao et al. (2025) obtain similarly high VBench Motion Smoothness scores of 0.965, 0.983, 0.981, and 0.976, respectively. However, their IAC scores differ substantially: 0.969 for CueInsert, 0.813 for Senorita, 0.156 for AnyV2V, and 0 for ICVE. This discrepancy reflects the limitation of generic smoothness metrics for insertion evaluation. AnyV2V and ICVE frequently fail to preserve a valid inserted object, while Senorita can produce temporally unstable insertions whose object appearance gradually degrades or disappears in later frames. Despite these failures, their global motion can remain temporally smooth and therefore receive high Motion Smoothness scores. In contrast, IAC explicitly measures insertion-specific object action and is more sensitive to whether meaningful foreground motion is actually maintained over time.

To further distinguish object deformation or articulated motion from prescribed trajectory translation, we remove the global displacement induced by the planned object center. For consecutive frames, let $F_t(\mathbf{x})$ denote optical flow within the object mask M_t, and let $\Delta\mathbf{c}_t = \mathbf{c}_{t+1} - \mathbf{c}_t$ denote the planned center displacement. We compute the residual flow as

$$F_t^{\text{res}}(\mathbf{x}) = F_t(\mathbf{x}) - \Delta\mathbf{c}_t, \qquad \mathbf{x} \in M_t, \tag{13}$$

such that rigid trajectory translation is suppressed while local object motion is retained. This enables IAC to characterize non-rigid or articulated object motion rather than merely rewarding translation along the prescribed trajectory.

D Additional Results

We provide additional qualitative results on two cases: inserting a skydiver (kite-walk) into the *Rhino* video 8 and inserting a bear into the *BMX-bumps* video 9. These examples further demonstrate CueInsert's ability to preserve object fidelity while producing coherent motion and plausible scene interactions.

E Limitations

CueInsert depends on reliable scene understanding, anchor tracking, and semantic proposal generation. Performance may degrade under severe occlusion, strong camera motion, ambiguous spatial relations, or complex articulated motion. The current relation predicates also cover only a predefined set of interaction intents, while final rendering quality remains bounded by the pretrained video generator.



Figure 8: **Qualitative comparison on Bear→BMX-bumps.** CueInsert better preserves object fidelity and temporal consistency.

F Broader Impact

CueInsert may support controllable video editing, simulation, and data augmentation, but could also facilitate misleading visual manipulation. Responsible use, disclosure of synthetic content, and appropriate safeguards are therefore important for real-world deployment.

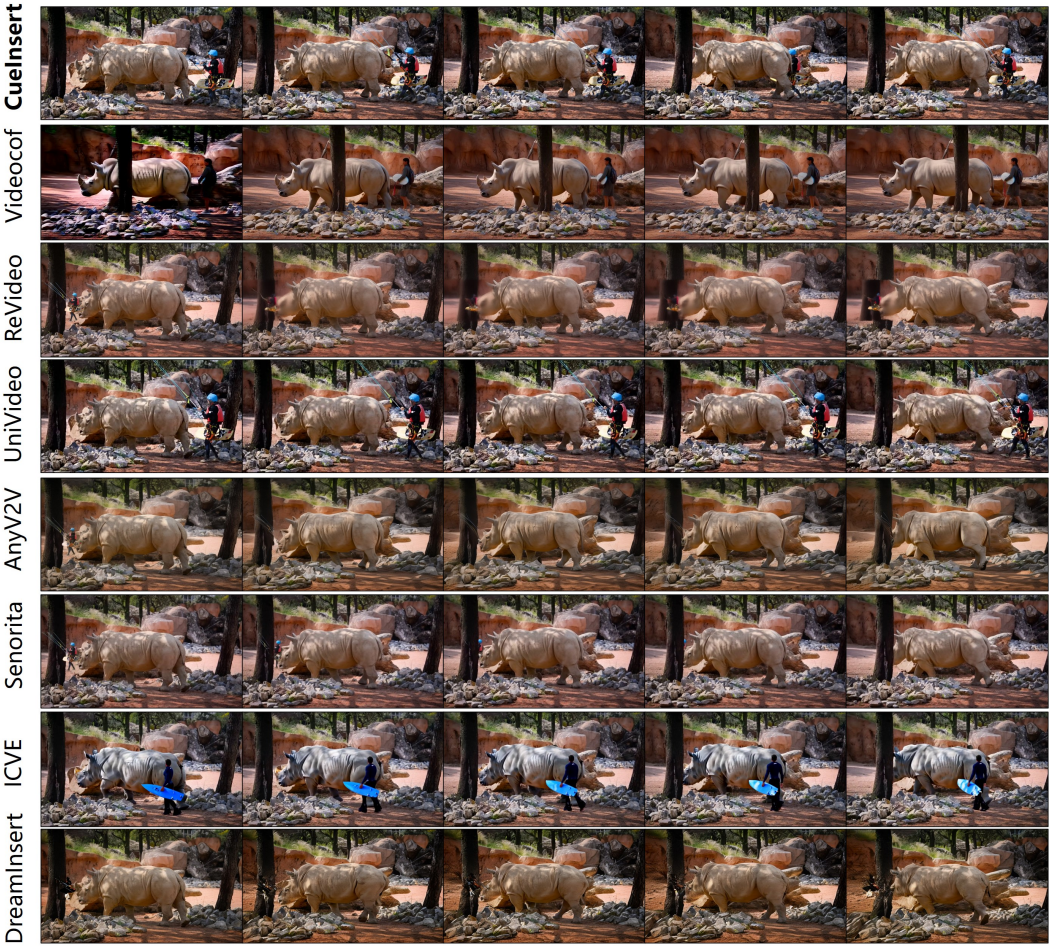

Figure 9: **Qualitative comparison on Kite-walk→Rhino.** CueInsert preserves the inserted object while maintaining coherent motion and plausible scene interaction.