VIDEO REVERSAL CURSE: DIAGNOSING AND ALIGNING TEMPORAL REASONING IN DIFFUSION VLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Language Models can exhibit the Reversal Curse: after learning that event A precedes event B, they may fail to infer that event B follows event A. We show that video-language models exhibit the analogous **Video Reversal Curse**, combining directional bias with weak reliance on temporal order: they answer cause-to-effect questions more reliably than paired effect-to-cause questions and may ignore frame reversals or shuffles. Although dVLMs provide bidirectional attention and iterative denoising, these capabilities alone do not ensure symmetric temporal reasoning. We propose **ReVID**, which combines Temporal-Symmetric GRPO, training on paired forward/backward questions with rewards for balanced accuracy and temporal reliance, and Temporal Visual Contrastive Decoding, which contrasts original and shuffled-video logits at inference. We introduce **RevVideo**, a 3,334-question paired benchmark, and the Temporal Symmetry Score. ReVID improves RevVideo accuracy by 8.37 points over its backbone, reduces the forward/backward gap to 1.78 points, and achieves the best TSS of 40.56, while improving dVLM performance on TempCompass and TVBench.

1 INTRODUCTION

Large Language Models (LLMs) suffer from the Reversal Curse: after learning that Event A is followed by Event B, they may fail to infer that Event B was preceded by Event A Berglund et al. (2024). In video understanding, this limitation is more severe because temporal reasoning requires inferring both what happens next and what happened before. We refer to this manifestation as the *Video Reversal Curse*.

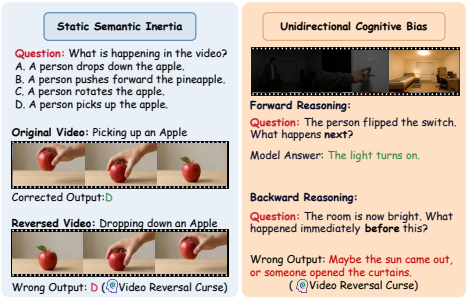


Figure 1: **Video Reversal Curse.** ARMs rely on static cues under frame reversal and show asymmetric cause→effect versus effect→cause reasoning.

Recent autoregressive video models improve temporal understanding through stronger video encoders, post-training, and contrastive decoding Feng et al. (2025); Wang et al. (2025b); Wu et al. (2025a). However, their strict left-to-right generation makes backward attribution difficult, since earlier representations cannot be revised using later visual evidence. This limitation produces two characteristic failure modes as illustrated in the Figure 1. *Static Semantic Inertia* occurs when reversing the chronological order of frames leaves the prediction unchanged, indicating reliance on static object-level cues rather than causal trajectories. *Unidirectional Cognitive Bias* occurs when models answer cause→effect questions more reliably than the corresponding effect→cause queries, revealing an imbalance between the two directions of temporal reasoning.

dVLMs provide a natural alternative because masked denoising enables bidirectional attention and iterative refinement over visual and textual context Wang et al. (2025a). But off-the-shelf dVLMs still exhibit forward/backward performance gaps, showing that bidirectional attention provides the capacity for temporal symmetry but does not explicitly enforce it during training or decoding.

To address this gap, we propose ReVID, a dVLM-based framework for mitigating the Video Reversal Curse. ReVID combines Temporal-Symmetric GRPO (TS-GRPO), which trains on paired

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.

forward/backward questions from the same video with rewards for balanced temporal reasoning, and Temporal Visual Contrastive Decoding (TVCD), which contrasts original and shuffled-frame logits to suppress order-invariant visual evidence during inference.

Our contributions are summarized as follows:

• We propose ReVID, the first dVLM-based framework designed to address the Video Reversal Curse. Its two core components, TS-GRPO and TVCD, target training-time directional bias and inference-time static cue reliance.

• We introduce RevVideo, a benchmark with 3,334 paired forward/backward questions across four temporal reasoning categories, together with the Temporal Symmetry Score for measuring bidirectional consistency and directional balance.

• Experiments on RevVideo, TempCompass, and TVBench show that ReVID consistently improves temporal reasoning. It improves RevVideo Overall accuracy by 8.37 points over its backbone, reduces the Forward/Backward gap to 1.78 points, and achieves the highest TSS of 40.56 points.

2 Related Work

2.1 Large Language Diffusion Models

Diffusion models have recently been extended from continuous generation to discrete text generation through masked diffusion. LLaDA Nie et al. (2025) demonstrates that diffusion language models can scale to billion-parameter settings and compete with autoregressive models, while DREAM Ye et al. (2025) and MMaDA Yang et al. (2025) improve instruction following and multimodal generation. Unlike autoregressive models, diffusion language models use bidirectional attention, allowing each token to access the full sequence and jointly model forward and backward dependencies. Recent studies further explore reinforcement learning Zhao et al. (2025); Wang et al. (2025c); Zhang et al. (2026), inference acceleration Wu et al. (2025b), and remasking-based reliability improvements Hemmat et al. (2026); Zhai et al. (2026).

2.2 Video Large Language Models

Video MLLMs must interpret dynamic visual content, track state changes, and reason over temporally ordered events Li et al. (2024a;b); Liu et al. (2023). Recent models have improved video understanding through long-context video encoding, spatiotemporal feature extraction, and audio-visual integration Shen et al. (2025); Shu et al. (2024); Xu et al. (2024); Zhang et al. (2024). However, most video MLLMs remain autoregressive, processing visual and textual tokens in a fixed left-to-right order. This design can induce static semantic inertia Liu et al. (2024); Cores et al. (2025) and temporal hallucination Li et al. (2025a); Wu et al. (2025a), particularly when questions require backward attribution from observed outcomes to preceding causes. Video-R1 Feng et al. (2025) and Time-R1 Wang et al. (2025b) use R1-style GRPO to strengthen temporal reasoning, while contrastive decoding methods perturb visual evidence at inference time. These approaches improve temporal grounding but do not remove the causal decoding constraint of ARMs.

3 Video Reversal Curse in dVLMs

3.1 Failure Modes Analysis

The Video Reversal Curse describes a failure of temporal reasoning in which MLLMs rely on superficial frame-level semantics instead of causal dynamics. It has two main manifestations.

Static Semantic Inertia. As shown in Figure 1 (left), reversing the frames of "Picking up an Apple" produces the opposite event "Dropping down an Apple", yet a traditional ARM may retain the original prediction. This behavior indicates reliance on static object cues rather than causal trajectories and motivates frame-shuffling contrast as an explicit diagnostic.

Unidirectional Cognitive Bias. As shown in Figure 1 (right), ARMs often answer cause-to-effect questions more reliably than paired effect-to-cause questions. This asymmetry stems from left-to-right causal attention, which prevents earlier-state representations from using later visual or textual evidence during backward inference.

3.2 Why dVLMs Are the Principled Cure

Both failure modes arise from the left-to-right generation constraint of ARMs. dVLMs provide two complementary architectural properties that better support bidirectional temporal reasoning, as illustrated in Figure 2.

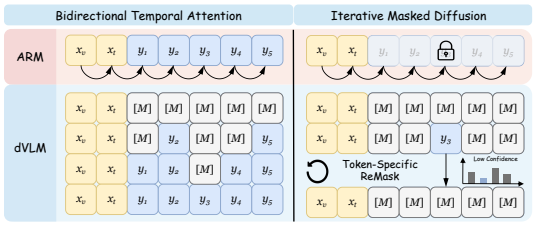


Figure 2: **Architectural Comparison of ARMs and dVLMs.** dVLMs use bidirectional temporal attention and iterative masked refinement.

Bidirectional Temporal Attention. At each denoising step, every token attends to the full sequence, allowing representations of earlier states to incorporate later temporal evidence. This removes the architectural bottleneck that makes backward inference difficult in ARMs and limits their ability to exploit future context.

Iterative Masked Diffusion. Masked diffusion repeatedly updates predictions across denoising steps, allowing uncertain early outputs to be revised as more temporal context becomes available. This iterative refinement supports the reconstruction of causal antecedents and provides a correction mechanism unavailable in standard autoregressive decoding during generation.

LLaDA Nie et al. (2025) shows that diffusion language models can mitigate the reversal curse in text. We use this result as a sanity check and motivation for extending the same structural advantage to video understanding through ReVID.

4 Method

In this section, we present a comprehensive framework, ReVID, designed to eradicate the Video Reversal Curse in dVLMs. Our approach acknowledges that the reversal curse stems from both a training-time directional bias and an inference-time over-reliance on static semantic priors. To this end, we first establish the diffusion-based generative paradigm of LLaDA-V. Subsequently, we introduce TS-GRPO, a reinforcement learning strategy that enforces bidirectional causal consistency. Finally, we detail TVCD, an inference-time mechanism that amplifies temporally dependent logits while suppressing static hallucinations. The overall framework is illustrated in Figure 3.

4.1 Preliminaries: Discrete Diffusion VLMs

Our framework is built upon the discrete diffusion paradigm, specifically leveraging the LLaDA-V architecture. Unlike autoregressive (AR) models that are constrained by a fixed left-to-right generation order, dVLMs utilize bidirectional attention, making them well suited for modeling complex temporal dynamics in which future frames can inform the interpretation of past events and improve temporally grounded video understanding.

LLaDA-V models the joint distribution of a sequence $\mathbf{x}_0$ by iteratively refining a fully masked sequence $\mathbf{x}_T = [\text{MASK}]^L$. The forward transition, or the masking process, at any continuous time step $t \in [0, 1]$ is formulated as:

$$q(x_t \mid x_0) = \prod_{i=1}^{n} \text{Cat}(x_t^i; (1-t)\delta_{x_0^i} + t\delta_{[\text{MASK}]}), \tag{1}$$

where δ is the Kronecker delta representing a point mass. The model is trained to reverse this process by maximizing the evidence lower bound (ELBO). The optimization objective is defined as the expectation over all possible masking levels:

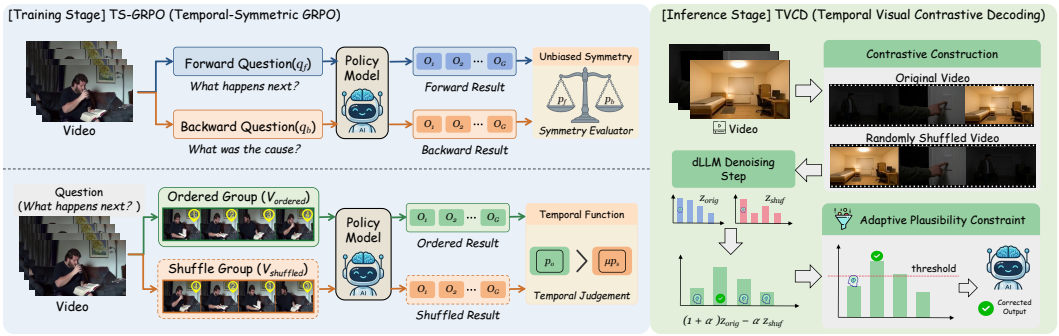


Figure 3: **ReVID framework.** TS-GRPO trains on paired forward/backward questions with symmetry and temporal reliance rewards. TVCD contrasts original and shuffled-video logits during dVLM decoding under an adaptive plausibility constraint.

$$\mathcal{J}_{\text{full}}(\theta) = \mathbb{E}_{t,\, q(x_t|x_0)} \left[\frac{1}{t|x_0|} \sum_{i \in \mathcal{M}_t} \log p_\theta(x_0^i \mid x_t, Q, v) \right], \tag{2}$$

where $\mathcal{M}_t = \{i : x_t^i = [\text{MASK}]\}$, Q denotes the linguistic prompt, and v represents the visual context. During inference, the model recovers $\mathbf{x}_0$ from $\mathbf{x}_T$ through a decoding trajectory $\tau = (\mathbf{x}_T, \mathbf{x}_{T-1}, \ldots, \mathbf{x}_0)$, iteratively unmasking tokens based on global context.

4.2 Temporal-Symmetric GRPO

To address the unidirectional cognitive bias inherent in MLLMs, we propose Temporal-Symmetric GRPO. Standard Group Relative Policy Optimization (GRPO) focuses on maximizing the relative advantage of responses within a group but fails to account for the temporal symmetry required for deep video understanding. TS-GRPO introduces a dual-reward mechanism that forces the model to internalize causal relationships as bidirectional logical links.

Question-Level Contrastive Trajectories. Our training strategy uses question-level contrast to couple two complementary temporal queries on the same video. For a video sequence v depicting a causal chain from Event A to Event B, the model should possess the ability to perform both forward prediction and backward attribution. We define a dual-task set: a Forward Question q_f (e.g., "What is the immediate consequence of Event A?") and a Backward Question q_b (e.g., "What was the necessary antecedent for Event B?").

Following *TracerRL*, we sample a group of G independent decoding trajectories $\{\tau_i\}_{i=1}^G$ for both question types. By evaluating the model across the entire trajectory rather than a single terminal state, we ensure that the temporal reasoning is consistent throughout the iterative refinement process.

Symmetric Reward Formulation. The final reward R_i assigned to each trajectory is a multi-faceted signal designed to decouple temporal logic from static language priors.

Unbiased Symmetry Reward. This component targets the Reversal Curse by penalizing the accuracy gap between forward and backward reasoning. Let p_f and p_b denote the group-level success rates for q_f and q_b, respectively. We define the unbiased symmetry reward r_{ub} as:

$$r_{ub} = \begin{cases} \eta, & \text{if } |p_f - p_b| < \delta \\ \eta \cdot \exp(-\frac{|p_f - p_b|}{\sigma}), & \text{otherwise} \end{cases}, \tag{3}$$

where δ is a tolerance threshold and σ is a temperature parameter controlling the penalty sharpness. This reward favors trajectories whose forward and backward success rates remain balanced, encouraging the model to improve bidirectional temporal reasoning and cure the directional bias.

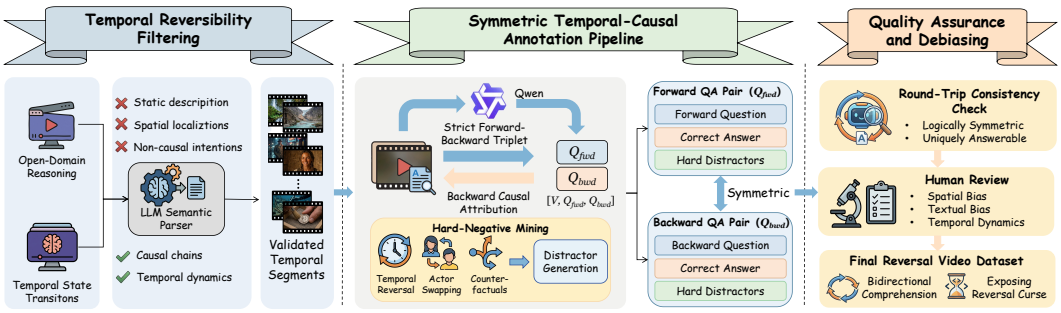


Figure 4: **RevVideo dataset construction pipeline.** It contains three stages: *Temporal Reversibility Filtering* selects videos with explicit causal dynamics; *Symmetric Temporal-Causal Annotation* generates paired forward/backward QA with hard negatives; *Quality Assurance and Debiasing* verifies logical symmetry and reduces shortcut biases.

Temporal Reliance Reward. To verify that the model's response is grounded in the video's temporal structure rather than static frame semantics, we utilize a frame-shuffling contrast. We compare the accuracy achieved using the ordered video frame sequence p with the accuracy $\tilde{p}$ obtained using a shuffled frame sequence $v' = \text{Shuffle}(v)$. The temporal reliance reward r_t is defined as:

$$r_t = \alpha \cdot \mathbb{I}(p_o > \mu \cdot p_s), \tag{4}$$

where $\alpha = 0.3$ and $\mu = 0.8$. This ensures that the model is only rewarded when its success is demonstrably dependent on the correct chronological order of information.

Policy Update with Trajectory Shrinkage. In the context of DLLMs, optimizing every single diffusion step is computationally prohibitive. Therefore, we utilize Trajectory Shrinkage with a parameter s, which aggregates s neighboring steps into a single update block. The policy π_θ is optimized using a clipped surrogate objective:

$$\begin{aligned}\mathcal{J}_{\text{TS-GRPO}}(\theta) = \mathbb{E}\Bigg[\frac{1}{G}\sum_{i=1}^{G}\sum_{k=1}^{|\tau_i^s|}\min\Bigg(&\frac{\pi_\theta(\mathbf{o}_k \mid \mathbf{x}_{k+1})}{\pi_{\text{old}}(\mathbf{o}_k \mid \mathbf{x}_{k+1})}A_i,\\ &\text{clip}\left(\frac{\pi_\theta(\mathbf{o}_k \mid \mathbf{x}_{k+1})}{\pi_{\text{old}}(\mathbf{o}_k \mid \mathbf{x}_{k+1})}, 1-\epsilon, 1+\epsilon\right)A_i\Bigg)\Bigg]\\ &- \beta D_{KL}(\pi_\theta \| \pi_{\text{ref}}),\end{aligned} \tag{5}$$

where A_i denotes the advantage computed from the standardized total reward $R_i = r_i + r_t + r_{ub}$, and $\mathbf{o}_k$ is the set of tokens unmasked in the k-th update block. This objective encourages the policy to preserve temporal evidence throughout the masked generation trajectory, rather than optimizing only the final decoded response after denoising.

4.3 Temporal Visual Contrastive Decoding

To mitigate the Video Reversal Curse during inference, we introduce TVCD. Unlike traditional Visual Contrastive Decoding (VCD) that injects noise into static images, TVCD shifts the perturbation to the temporal domain to isolate and amplify features genuinely dependent on chronological logic.

Contrastive View Construction. The core intuition of TVCD is to subtract pseudo-semantics that remain stable after temporal disruption. Given an input video v and a text prompt x, we uniformly sample a sequence of frames to obtain the original video representation. We then construct a temporally perturbed counterpart, v', by randomly shuffling the frame order:

$$v' = \text{Shuffle}(v). \tag{6}$$

Both the original sequence v and the shuffled sequence v' use the same text prompt x during inference, ensuring that the contrast isolates information associated with temporal order.

Temporal Contrastive Logits in dVLMs. At each decoding step t, LLaDA-V computes logits for the full sequence. Let $\mathbf{z}_{\text{orig}}^{(t)}$ and $\mathbf{z}_{\text{shuf}}^{(t)}$ denote logits conditioned on v and v'. We formulate the temporal contrastive logits as:

$$\mathbf{z}_{\text{vcd}}^{(t)} = (1+\alpha)\mathbf{z}_{\text{orig}}^{(t)} - \alpha\mathbf{z}_{\text{shuf}}^{(t)}, \tag{7}$$

where $\alpha > 0$ is a hyper-parameter. It controls the contrastive strength. Eq. 7 penalizes tokens that receive high confidence scores under both the correct and the shuffled temporal conditions, thereby suppressing static semantic inertia and highlighting precise temporal causality.

Adaptive Plausibility Constraint. To prevent the contrastive penalty from excessively distorting the distribution and hallucinations, we incorporate an Adaptive Plausibility Constraint. We define a dynamic threshold $\gamma^{(t)}$ based on the maximum logit of the original distribution at step t:

$$\gamma^{(t)} = \max_j z_{\text{orig},j}^{(t)} + \log\beta, \tag{8}$$

where $\beta \in (0, 1]$ serves as the plausibility parameter. We then truncate the logits, restricting the output space to only tokens deemed sufficiently plausible by the original model:

$$\tilde{z}_j^{(t)} = \begin{cases} z_{\text{vcd},j}^{(t)}, & \text{if } z_{\text{orig},j}^{(t)} \geq \gamma^{(t)} \\ -\infty, & \text{otherwise.} \end{cases} \tag{9}$$

Iterative Masked Decoding Refinement. Finally, the constrained probability distribution for the current step is obtained via the softmax function:

$$p_{\text{vcd}}^{(t)}(y_j \mid v, v', x) = \text{softmax}\left(\tilde{\mathbf{z}}^{(t)}\right)_j. \tag{10}$$

In LLaDA-V, this entire contrastive process is executed over the full sequence logits at every diffusion step. The resulting distribution $p_{\text{vcd}}^{(t)}$ is immediately utilized to perform the mask-fill update. By iterating this process throughout the masked decoding phase, the model robustly reconstructs causal antecedents, lifting the Video Reversal Curse.

5 Dataset Construction

We construct a new dataset, named *RevVideo*, to evaluate the Video Reversal Curse across four categories: behavior order, operation series, dynamic evolving, and temporal reasoning. Each video is paired with a forward question Q_{fwd} and a backward question Q_{bwd}, forming a symmetric triplet $[V, Q_{fwd}, Q_{bwd}]$ as shown in Figure 4. The benchmark is designed to reduce shortcuts from static frames or language priors and to test whether models genuinely rely on temporal evidence.

5.1 Data Collection and Reversibility Filtering

We curate video-text pairs with rich temporal dynamics from open-domain video trajectories and temporal reasoning datasets, including STAR Wu et al. (2024), Charades-STA Gao et al. (2017), Time-R1 Wang et al. (2025b), Video-R1 Feng et al. (2025), and MoVQA Zhang et al. (2025). To ensure temporal reversibility, we use an LLM semantic parser to analyze the original questions and CoT rationales, filtering out instances dominated by static descriptions, spatial localization, or intentions without clear chronological state changes. Only segments with explicit causal chains or sequential state transitions are retained for QA generation.

5.2 Symmetric Temporal Annotation Pipeline

To construct video reversed QA pairs, we use an advanced MLLM as a directed causal reasoning engine. For each temporally validated video segment, the MLLM extracts the key state transition from the video and its context, then synthesizes a strict Forward-Backward triplet (V, Q_{fwd}, Q_{bwd}). The forward question asks the model to predict the consequence or sequential successor, while the backward question requires causal attribution to the antecedent state or prerequisite action. To prevent text-only guessing, we apply hard-negative mining by restricting distractors to plausible in-video entities and actions, including reversed temporal sequences, swapped actors, and counterfactual choices.

Table 1: **RevVideo Bench statistics in four categories.** The dataset contains 1,479 videos and 1,667 symmetric forward and backward QA pairs, totaling 3,334 questions.

Category	Video Num	Fwd/Bwd Num
Behavior Order	75	77
Operation Series	158	301
Dynamic Evolving	45	45
Temporal Reasoning	1201	1244
Total	**1479**	**1667**

5.3 Quality Assurance and Debiasing

We use a two-stage quality assurance protocol to preserve the reversal constraint. First, an auxiliary evaluator performs round-trip consistency checks to verify that each Q_{fwd} and Q_{bwd} pair is logically symmetric, unambiguous, and uniquely answerable through temporal grounding. Second, human annotators review sampled triplets for residual spatial shortcuts, textual biases, and invalid temporal dynamics, ensuring that neither direction can be solved from a single static frame or open-world language priors across the two temporal directions.

5.4 Benchmark and Evaluation Framework

RevVideo contains 49k training samples and a balanced evaluation benchmark with 1,479 videos, 1,667 forward questions, and 1,667 backward questions. To prevent leakage, we split training and evaluation data at the video-source level before annotation and apply semantic filtering to remove near-duplicate QA pairs. Since standard accuracy can hide a forward/backward imbalance, we introduce the Temporal Symmetry Score (TSS) to measure instance-level bidirectional consistency. We denote the benchmark by the following set of paired video-question instances:

$$\mathcal{D} = \{(V_i, Q_{fwd}^{(i)}, Q_{bwd}^{(i)})\}_{i=1}^{N}. \tag{11}$$

For a given instance i, we define the correctness indicator functions for the forward and backward predictions to separately record whether each temporal direction is answered correctly:

$$c_{fwd}^{(i)} = \mathbb{I}(\hat{y}_{fwd}^{(i)} = y_{fwd}^{(i)}), \tag{12}$$

$$c_{bwd}^{(i)} = \mathbb{I}(\hat{y}_{bwd}^{(i)} = y_{bwd}^{(i)}), \tag{13}$$

where $\hat{y}$ represents the model's predicted option and y denotes the ground truth. The TSS is then mathematically formulated as the joint probability of bidirectional success:

$$TSS = \frac{1}{N} \sum_{i=1}^{N} \left(c_{fwd}^{(i)} \times c_{bwd}^{(i)} \right). \tag{14}$$

By applying this strict conjunction to the paired queries, the TSS credits an instance if and only if it successfully comprehends both the causal antecedent and its subsequent consequence, thereby explicitly exposing its inability to form unified bidirectional spatiotemporal representations.

6 Experiments

6.1 Experimental Setup

Table 2: **Performance on RevVideo.** All values are in %. TSS denotes Temporal Symmetry Score.

Method	Para.	Behav.	Oper.	Dyn.	Temp.	Fwd ↑	Bwd ↑	Overall ↑	TSS ↑
TCD-Video-LLaVA	ARM	54.55	11.30	46.67	61.25	53.63	49.43	51.53	33.89
Video-LLaVA-7B	ARM	55.19	10.96	47.78	61.54	53.87	49.61	51.74	34.19
LLaVA-NeXT-Video	ARM	62.34	8.64	55.56	63.30	52.97	53.39	53.18	36.35
VideoLLaMA2 7B	ARM	55.84	22.76	**73.33**	64.23	56.75	56.45	56.60	37.67
MMaDA	dVLM	40.00	12.79	50.00	46.18	39.64	40.66	40.15	20.84
LaViDa	dVLM	52.22	16.11	53.25	64.43	57.41	52.31	54.86	37.31
LLaDA-V	dVLM	57.78	22.09	54.55	56.95	51.89	49.25	50.57	30.53
LaViDa-ReVID (Ours)	dVLM	61.83	21.47	64.95	**67.88**	**60.74**	56.93	58.84	**40.56**
LLaDA-V-ReVID (Ours)	dVLM	**69.12**	**23.25**	71.79	64.50	59.83	**58.05**	**58.94**	39.03

Table 3: **Performance on TempCompass Event Order.**

Method	Paradigm	Event Order (%) ↑
LLaVA-1.5	ARM	34.4
SPHINX-v2	ARM	36.4
Qwen-VL-Chat	ARM	40.7
LLaMA-VID	ARM	41.4
Video-ChatGPT	Video-ARM	37.1
VideoChat2	Video-ARM	47.0
Video-LLaMA	Video-ARM	32.8
MMaDA	dVLM	33.44
LaViDa	dVLM	44.04
LLaDA-V	dVLM	54.64
ReVID (Ours)	dVLM	**66.23**

Table 4: **Performance comparison on TVBench across ARMs and dVLMs.** ReVID achieves the best results among dVLMs. All numeric values are in %.

Method	Para.	Act. Seq.↑	Ego. Seq.↑	Scene Trans.↑	Overall↑
VideoChat2	ARM	56.8	24.5	56.2	45.83
ST-LLM	ARM	51.7	24.0	54.4	43.37
Qwen2-VL 7B	ARM	63.8	22.5	75.7	54.00
VideoGPT	ARM	61.3	36.9	69.2	55.80
MMaDA	dVLM	46.91	21.50	40.00	39.17
LaViDa	dVLM	54.00	15.50	57.84	45.50
LLaDA-V	dVLM	64.53	26.00	63.78	54.99
ReVID (Ours)	**dVLM**	**65.90**	**32.00**	**69.73**	**58.52**

Datasets. We evaluate our framework on the following benchmarks. TVBench Cores et al. (2025) is a fine-grained temporal benchmark covering Action Sequence, Egocentric Sequence, and Scene Transition, totaling 822 samples. TempCompass Liu et al. (2024) evaluates temporal understanding with a focus on Event Order reasoning (302 samples). RevVideo comprises 3,334 samples across Behavior Order, Operation Series, Dynamic Evolving, and Temporal Reasoning subsets, with balanced Forward and Backward question splits designed to directly probe the Video Reversal Curse.

Baselines. We compare against dVLM-based multimodal models: MMaDA Yang et al. (2025), LaViDa Li et al. (2025b), and LLaDA-V You et al. (2025), which serves as our direct backbone. All baselines are evaluated under standard inference settings without temporal-specific augmentation.

Implementation Details. ReVID integrates three complementary components on dVLMs: supervised fine-tuning on RevVideo training data (SFT), reinforcement learning via TS-GRPO, and Temporal Visual Contrastive Decoding at inference time.

6.2 Quantitative Results

The quantitative results across all three benchmarks consistently demonstrate the effectiveness of our framework in addressing the Video Reversal Curse. On TempCompass (Table 3), it reaches 66.23% Event Order accuracy, exceeding LLaDA-V by 11.59 points and MMaDA by 32.79 points. On TVBench (Table 4), it achieves 58.52% Overall accuracy, improving LLaDA-V by 3.53 points, with gains on Egocentric Sequence (+6.00) and Scene Transition (+5.95).

RevVideo (Table 2) directly measures directional asymmetry. LaViDa has a 5.10-point Forward/Backward gap, and LLaDA-V has a 2.64-point gap. The ReVID variants reduce this gap to 1.78 points (LLaDA-V-ReVID), while LaViDa-ReVID attains the best TSS (40.56) and LLaDA-V-

Table 5: **Ablation on RevVideo.** We remove training, inference, and reward components from ReVID on LLaDA-V. All numeric values are in %.

Method	Behav. ↑	Oper. ↑	Dyn. ↑	Temp. ↑	Fwd ↑	Bwd ↑	Overall ↑	TSS ↑
ReVID (Full)	**69.12**	23.25	**71.79**	**64.50**	**59.83**	**58.05**	**58.94**	**39.03**
w/o SFT	55.88	**24.75**	56.41	59.64	53.85	55.13	54.49	33.26
w/o TS-GRPO	54.41	14.00	56.41	55.53	49.29	49.93	49.61	29.20
w/o TVCD	60.29	13.75	66.67	52.61	47.15	48.15	47.65	25.07
w/o r_t	60.29	16.00	58.97	57.65	52.92	50.71	51.82	32.34
w/o r_{ub}	67.65	15.25	57.69	58.75	53.21	52.28	52.74	33.69

ReVID attains the best Overall accuracy (58.94%). This near-symmetric bidirectional performance indicates that ReVID improves bidirectional balance and substantially reduces the structural directional bias underlying the Reversal Curse.

6.3 Qualitative Analysis

Figure 5 compares forward/backward QA examples. Advanced baselines often answer Q_{fwd} correctly but fail on Q_{bwd}, which requires inferring preceding events. ReVID answers both directions correctly, suggesting that bidirectional optimization improves temporal grounding beyond forward-only reasoning. This consistent performance qualitatively proves that explicitly optimizing for bidirectional consistency allows ReVID to successfully overcome the Reversal Curse and acquire an unbiased understanding of temporal causality.

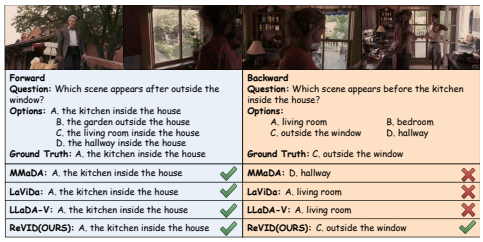


Figure 5: **Qualitative Comparison.**

6.4 Ablation Study

Component Removal. Table 5 shows that each component contributes to the full ReVID model. Removing SFT reduces overall accuracy from 58.94 to 54.49 and TSS from 39.03 to 33.26, indicating that task-specific temporal supervision is useful. Removing TS-GRPO causes a larger drop to 49.61 Overall and 29.20 TSS, showing the importance of bidirectional reward shaping. Removing TVCD gives the lowest TSS (25.07), suggesting that inference-time temporal contrast is critical for suppressing order-invariant evidence.

Reward Ablation. The last two rows isolate the two TS-GRPO rewards while keeping the remaining components fixed. Removing the Temporal Reliance Reward r_t lowers Overall accuracy to 51.82 and TSS to 32.34, while removing the Unbiased Symmetry Reward r_{ub} gives 52.74 Overall and 33.69 TSS. Both rewards are therefore necessary: r_t encourages dependence on chronological evidence, and r_{ub} penalizes forward/backward imbalance. More experiments and analyses are provided in the supplementary material.

7 Conclusion

We presented a comprehensive framework to cure the Video Reversal Curse in dVLMs, combining TS-GRPO for bidirectional reward shaping, TVCD for inference-time temporal amplification, and an Adaptive Plausibility Constraint to suppress hallucinations. To support evaluation, we introduced RevVideo and the TSS. Experiments on TVBench, TempCompass, and RevVideo demonstrate consistent state-of-the-art performance, with LaViDa-ReVID achieving the highest TSS of 40.56 and LLaDA-V-ReVID reducing the Forward/Backward accuracy gap to 1.78 while reaching 58.94% overall accuracy, confirming that the bidirectional nature of dVLMs, when properly trained and decoded, constitutes a principled and effective cure for the Video Reversal Curse.

AI USE STATEMENT

Generative AI tools were used only to assist with language editing, proofreading, and presentation refinement. No AI tools were used to generate data, fabricate results, conduct experiments, or make scientific decisions; all AI-assisted content was reviewed and verified by the authors, who take full responsibility for the final manuscript.

ETHICS STATEMENT

This work uses publicly available video datasets and does not involve human subjects, private information, or interventions. We follow the licenses and release practices of the source datasets and identify no additional ethical risks beyond those associated with the original data.

REFERENCES

Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. The reversal curse: Llms trained on "a is b" fail to learn "b is a", 2024. URL https://arxiv.org/abs/2309.12288.

Daniel Cores, Michael Dorkenwald, Manuel Mucientes, Cees G. M. Snoek, and Yuki M. Asano. Lost in Time: A New Temporal Benchmark for VideoLLMs, March 2025.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Benyou Wang, and Xiangyu Yue. Video-R1: Reinforcing Video Reasoning in MLLMs, March 2025.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query, 2017. URL https://arxiv.org/abs/1705.02101.

Arshia Hemmat, Philip Torr, Yongqiang Chen, and Junchi Yu. TDGNet: Hallucination Detection in Diffusion Language Models via Temporal Dynamic Graphs, February 2026.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer, 2024a. URL https://arxiv.org/abs/2408.03326.

Chaoyu Li, Eun Woo Im, and Pooyan Fazli. VidHalluc: Evaluating Temporal Hallucinations in Multimodal Large Language Models for Video Understanding, March 2025a.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models, 2024b. URL https://arxiv.org/abs/2407.07895.

Shufan Li, Konstantinos Kallidromitis, Hritik Bansal, Akash Gokul, Yusuke Kato, Kazuki Kozuka, Jason Kuen, Zhe Lin, Kai-Wei Chang, and Aditya Grover. LaViDa: A Large Diffusion Language Model for Multimodal Understanding, May 2025b.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning, December 2023.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. TempCompass: Do Video LLMs Really Understand Videos?, June 2024.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large Language Diffusion Models, February 2025.

Leqi Shen, Tao He, Guoqiang Gong, Fan Yang, Yifeng Zhang, Pengzhang Liu, Sicheng Zhao, and Guiguang Ding. LLaVA-MLB: Mitigating and Leveraging Attention Bias for Training-Free Video LLMs, March 2025.

Yan Shu, Zheng Liu, Peitian Zhang, Minghao Qin, Junjie Zhou, Zhengyang Liang, Tiejun Huang, and Bo Zhao. Video-xl: Extra-long vision language model for hour-scale video understanding, 2024. URL https://arxiv.org/abs/2409.14485.

Wen Wang, Bozhen Fang, Chenchen Jing, Yongliang Shen, Yangyi Shen, Qiuyu Wang, Hao Ouyang, Hao Chen, and Chunhua Shen. Time is a feature: Exploiting temporal dynamics in diffusion language models, 2025a. URL https://arxiv.org/abs/2508.09138.

Ye Wang, Ziheng Wang, Boshen Xu, Yang Du, Kejun Lin, Zihan Xiao, Zihao Yue, Jianzhong Ju, Liang Zhang, Dingyi Yang, Xiangnan Fang, Zewen He, Zhenbo Luo, Wenxuan Wang, Junqi Lin, Jian Luan, and Qin Jin. Time-R1: Post-Training Large Vision Language Model for Temporal Video Grounding, June 2025b.

Yinjie Wang, Ling Yang, Bowen Li, Ye Tian, Ke Shen, and Mengdi Wang. Revolutionizing Reinforcement Learning Framework for Diffusion Large Language Models, September 2025c.

Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. Star: A benchmark for situated reasoning in real-world videos, 2024. URL https://arxiv.org/abs/2405.09711.

Chang-Hsun Wu, Kai-Po Chang, Yu-Yang Sheng, Hung-Kai Chung, Kuei-Chun Wang, and Yu-Chiang Frank Wang. SEASON: Mitigating Temporal Hallucination in Video Large Language Models via Self-Diagnostic Contrastive Decoding, December 2025a.

Chengyue Wu, Hao Zhang, Shuchen Xue, Zhijian Liu, Shizhe Diao, Ligeng Zhu, Ping Luo, Song Han, and Enze Xie. Fast-dllm: Training-free acceleration of diffusion llm by enabling kv cache and parallel decoding, 2025b. URL https://arxiv.org/abs/2505.22618.

Guowei Xu, Peng Jin, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. LLaVA-CoT: Let Vision Language Models Reason Step-by-Step, November 2024.

Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. MMaDA: Multimodal Large Diffusion Language Models, May 2025.

Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. Dream 7b: Diffusion large language models, 2025. URL https://arxiv.org/abs/2508.15487.

Zebin You, Shen Nie, Xiaolu Zhang, Jun Hu, Jun Zhou, Zhiwu Lu, Ji-Rong Wen, and Chongxuan Li. LLaDA-V: Large Language Diffusion Models with Visual Instruction Tuning, May 2025.

Kevin Zhai, Sabbir Mollah, Zhenyi Wang, and Mubarak Shah. CORE: Context-Robust Remasking for Diffusion Language Models, February 2026.

Hongjie Zhang, Lu Dong, Yi Liu, Yifei Huang, Yali Wang, Limin Wang, and Yu Qiao. Lvbench: A benchmark for long-form video understanding with versatile multi-modal question answering, 2025. URL https://arxiv.org/abs/2312.04817.

Wenxuan Zhang, Lemeng Wu, Changsheng Zhao, Ernie Chang, Mingchen Zhuge, Zechun Liu, Andy Su, Hanxian Huang, Jun Chen, Chong Zhou, Raghuraman Krishnamoorthi, Vikas Chandra, Mohamed Elhoseiny, and Wei Wen. dTRPO: Trajectory Reduction in Policy Optimization of Diffusion Large Language Models, March 2026.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video Instruction Tuning With Synthetic Data, October 2024.

Siyan Zhao, Devaansh Gupta, Qinqing Zheng, and Aditya Grover. D1: Scaling Reasoning in Diffusion Large Language Models via Reinforcement Learning, June 2025.

A Dataset Leakage Prevention and Annotation Quality

A.1 Train/Evaluation Split Independence

To prevent leakage between the 49k training split and the 3,334-sample evaluation benchmark, we enforce strict source-level separation throughout the dataset construction pipeline. Training and

Table 7: End-to-end generation latency comparison between vanilla LLaDA-V and TVCD across model variants. All runs use identical hardware and decoding settings. Latency and throughput are reported in seconds.

Model	Vanilla Latency(s)	TVCD Latency(s)	Δ Latency(s)	Overhead	Throughput Vanilla/TVCD(s/sample)
Base	6.925	11.789	+4.864	1.702× (+70.2%)	0.144 / 0.085
SFT	6.915	11.874	+4.959	1.717× (+71.7%)	0.145 / 0.084
TS-GRPO	6.904	11.863	+4.959	1.718× (+71.8%)	0.145 / 0.084
SFT+TS-GRPO	6.893	11.772	+4.879	1.708× (+70.8%)	0.145 / 0.085
Average	**6.909**	**11.825**	**+4.916**	**1.711× (+71.1%)**	**0.145 / 0.085**

Table 8: Full ablation study on RevVideo. Each row corresponds to a specific combination of SFT, TS-GRPO (RL), and TVCD (INF). The Forward/Backward split and TSS measure bidirectional temporal symmetry.

SFT	RL	INF	B. Ord. (%)	O. Ser. (%)	D. Evol. (%)	T. Reas. (%)	Fwd (%)	Bwd (%)	Ovrl. (%)	TSS
×	×	×	57.78	22.09	54.55	56.95	51.89	49.25	50.57	30.53
✓	×	×	60.29	14.50	60.26	56.76	50.71	51.14	50.93	30.77
×	✓	×	48.53	23.25	55.13	57.91	53.77	51.57	52.67	30.41
×	×	✓	42.65	19.25	56.41	59.64	53.85	52.92	53.38	31.84
✓	✓	×	60.29	13.75	66.67	52.61	47.15	48.15	47.65	25.07
×	✓	✓	55.88	**24.75**	56.41	59.64	53.85	55.13	54.49	33.26
✓	×	✓	54.41	14.00	56.41	55.53	49.29	49.93	49.61	29.20
✓	✓	✓	**69.12**	23.25	**71.79**	**64.50**	**59.83**	**58.05**	**58.94**	**39.03**

evaluation samples are partitioned at the *video source* level before any annotation step, so no evaluation video is ever used to construct training samples. The LLM-based symmetric temporal-causal annotation pipeline is then applied independently to each partition, and a semantic deduplication pass removes any residual near-duplicate questions across splits.

A.2 Human Review Coverage and Annotation Quality

All annotations, including those initially drafted by the LLM-assisted pipeline, are independently verified and edited by trained human annotators. The LLM serves only as a drafting aid, not as the final arbiter of quality.

A fraction of 30% of the evaluation benchmark, corresponding to 500 sampled triplets, undergoes full human expert review. Annotators inspect each forward–backward triplet $(V, Q_{\text{fwd}}, Q_{\text{bwd}})$ for residual shortcut exploitability, logical symmetry, and temporal validity.

The metrics confirm that annotators actively reviewed and revised LLM-generated outputs, establishing genuine human oversight throughout the construction process. Taken together, source-level split separation, semantic deduplication, and human review show that RevVideo avoids the principal forms of dataset bias and leakage.

Table 6: Annotation quality metrics for the RevVideo evaluation benchmark.

Metric	Score
Human review coverage (%)	30
Inter-annotator agreement (Cohen's κ)	0.86
Annotations substantially modified post-LLM (%)	28.4

B Inference Overhead and Caching Analysis

To quantify the inference cost introduced by TVCD, we compare it with vanilla LLaDA-V under the same evaluation pipeline and decoding settings. We use 8 sampled frames, 128 denoising steps, and 128 generated tokens per sample. Table 7 summarizes the measured generation latency.

Latency. Vanilla LLaDA-V requires 6.91 s/sample on average, whereas TVCD requires 11.82 s/sample. This corresponds to a stable $1.71\times$ latency overhead, or about 71%, across all model variants. The overhead is expected: TVCD adds one contrastive forward pass at each denoising step, computes logits from a temporally shuffled view in parallel with the original branch, and combines the two distributions during decoding. The measured latency refers to the generation call itself; frame sampling and video decoding are shared preprocessing steps, so the full end-to-end overhead is slightly smaller.

Memory. TVCD adds no trainable parameters. It does, however, maintain a separate embedding stream and cache state for the contrastive branch during inference, so peak GPU memory is higher than for vanilla LLaDA-V. Our current implementation does not log peak memory explicitly, so we avoid reporting an unverified ratio here.

Caching. KV caching is used within each sample across denoising steps, but not across different videos. The cache is reset at the beginning of every generation call, so each video is decoded independently. When TVCD is enabled, the original and shuffled branches maintain separate cache states, which preserves correctness and avoids cross-sample leakage.

C More Qualitative Results

To further illustrate the Video Reversal Curse, we provide additional qualitative comparisons across complex physical collisions, sequential laboratory procedures, and fine-grained daily activities. Across these scenarios, standard diffusion-based MLLM baselines such as MMaDA, LaViDa, and LLaDA-V often succeed on the forward question but fail on the backward one. When forced to infer the antecedent, they fall back on fragmented semantic priors and produce hallucinatory answers rather than a true temporal explanation.

ReVID behaves differently. By optimizing temporal symmetry during training and suppressing static semantic hallucinations during inference, it produces consistent forward and backward answers without relying on unidirectional shortcuts. These cases complement the quantitative results and show that ReVID substantially mitigates the Video Reversal Curse.

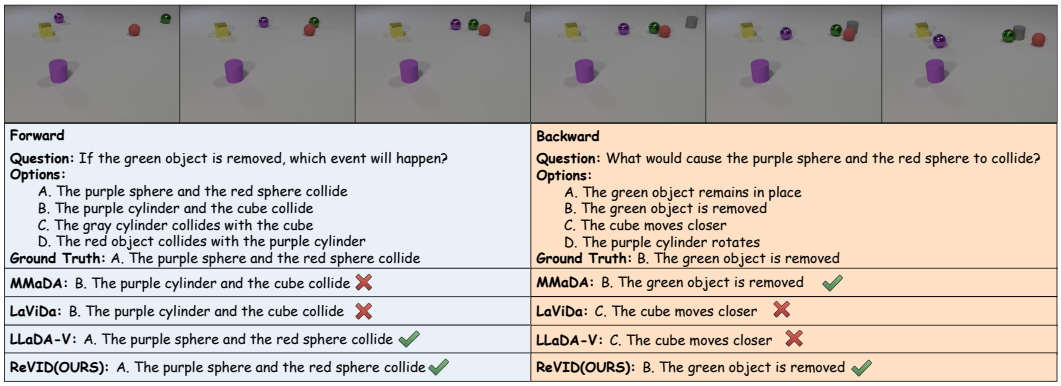
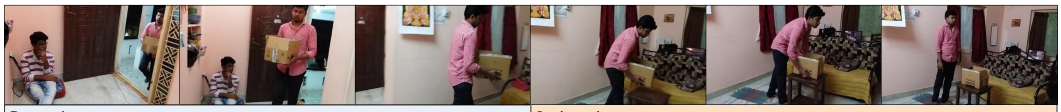


Figure 6: More qualitative results.

D More Ablation Analysis

D.1 Full Ablation Grid

The complete ablation grid is shown in Table 8. It confirms that the full ReVID configuration is strongest on every key metric and that no partial configuration matches its overall balance.



Forward	Backward
Question: What is the final action shown in the video? **Options:** A. The person in the red checkered shirt opens the box B. The person in the red checkered shirt leaves the room C. The person in the red checkered shirt stands next to the table with the box securely placed on it D. The person in the red checkered shirt sits on the sofa **Ground Truth:** C. The person in the red checkered shirt stands next to the table with the box securely placed on it	**Question:** What action directly precedes the person standing next to the table with the box? **Options:** A. The person carries the box into the room B. The person opens the box C. The person places the box on the table D. The person sits on the sofa **Ground Truth:** C. The person places the box on the table
MMaDA: A. The person in the red checkered shirt opens the box ✗	**MMaDA:** B. The person opens the box ✗
LaViDa: A. The person in the red checkered shirt opens the box ✗	**LaViDa:** A. The person carries the box into the room ✗
LLaDA-V: C. The person in the red checkered shirt stands next to the table with the box securely placed on it ✓	**LLaDA-V:** A. The person carries the box into the room ✗
ReVID(OURS): C. The person in the red checkered shirt stands next to the table with the box securely placed on it ✓	**ReVID(OURS):** C. The person places the box on the table ✓

Figure 7: More qualitative results.



Forward	Backward
Question: After pouring the conical flask, what step did this person perform? **Options:** A. take up the test tube B. put down the conical flask C. screw the iron clamp D. take up the horn tube E. take up the conical flask F. insert the horn tube **Ground Truth:** B. put down the conical flask	**Question:** What steps did this person take before putting down the conical flask? **Options:** A. take up the test tube B. pour the conical flask C. screw the iron clamp D. take up the horn tube E. take up the conical flask F. insert the horn tube **Ground Truth:** B. pour the conical flask
MMaDA: A. take up the test tube ✗	**MMaDA:** A. take up the test tube ✗
LaViDa: B. put down the conical flask ✓	**LaViDa:** E. take up the conical flask ✗
LLaDA-V: C. screw the iron clamp ✗	**LLaDA-V:** B. pour the conical flask ✓
ReVID(OURS): B. put down the conical flask ✓	**ReVID(OURS):** B. pour the conical flask ✓

Figure 8: More qualitative results.



Forward	Backward
Question: What did the person do after opening the closet/cabinet? **Options:** A. Washed the hands B. Took the dish C. Put on the shoes D. Closed the door **Ground Truth:** B. Took the dish	**Question:** What did the person do before taking the dish? **Options:** A. Poured the soup B. Put on the gloves C. Opened the closet/cabinet D. Washed the hands **Ground Truth:** C. Opened the closet/cabinet
MMaDA: B. Took the dish ✓	**MMaDA:** B. Put on the gloves ✗
LaViDa: D. Closed the door ✗	**LaViDa:** D. Washed the hands ✗
LLaDA-V: B. Took the dish ✓	**LLaDA-V:** D. Washed the hands ✗
ReVID(OURS): B. Took the dish ✓	**ReVID(OURS):** C. Opened the closet/cabinet ✓

Figure 9: More qualitative results.

Table 9: Reward component ablation on RevVideo. Removing r_t removes the Temporal Reliance Reward; removing r_{ub} removes the Unbiased Symmetry Reward.

Method	B. Ord. (%)	O. Ser. (%)	D. Evol. (%)	T. Reas. (%)	Fwd (%)	Bwd (%)	Ovrl. (%)	TSS
ReVID (Full)	**69.12**	**23.25**	**71.79**	**64.50**	**59.83**	**58.05**	**58.94**	**39.03**
w/o r_t	60.29	16.00	58.97	57.65	52.92	50.71	51.82	32.34
w/o r_{ub}	67.65	15.25	57.69	58.75	53.21	52.28	52.74	33.69

Table 10: Category-level breakdown for key configurations. Behavior Order and Dynamic Evolving are the most causally demanding categories; Temporal Reasoning is the most data-abundant.

Config	B. Ord. (%)	O. Ser. (%)	D. Evol. (%)	T. Reas. (%)	Ovrl. (%)	TSS
Backbone	57.78	22.09	54.55	56.95	50.57	30.53
SFT only	60.29	14.50	60.26	56.76	50.93	30.77
TS-GRPO only	48.53	23.25	55.13	57.91	52.67	30.41
TVCD only	42.65	19.25	56.41	59.64	53.38	31.84
SFT+TS-GRPO	60.29	13.75	66.67	52.61	47.65	25.07
TS-GRPO+TVCD	55.88	24.75	56.41	59.64	54.49	33.26
SFT+TVCD	54.41	14.00	56.41	55.53	49.61	29.20
Full (ReVID)	**69.12**	**23.25**	**71.79**	**64.50**	**58.94**	**39.03**

D.2 Component-wise Contribution Analysis

We summarize the single-component gains over the backbone in Table 11. No single component alone produces a dominant gain in Overall accuracy: SFT adds 0.36 points, TS-GRPO adds 2.10 points, and TVCD adds 2.81 points. The full model reaches 8.37 points, which is clearly synergistic and also the most balanced configuration in the forward/backward split.

D.3 Reconciling Component Interactions

Table 10 provides a category-level breakdown to further illuminate the complementary roles of each component. Three patterns stand out. First, SFT+TS-GRPO without TVCD yields the lowest Overall and TSS among all non-trivial configurations, which suggests that training-time adaptation alone can overfit the symmetry reward. Second, TS-GRPO+TVCD recovers strongly, showing that TVCD is essential to translate bidirectional training gains into robust inference-time behavior. Third, SFT without TS-GRPO consistently suppresses Operation Series performance, which indicates that passive imitation is not enough for the most temporally complex sequences. Overall, the full model is the only configuration that performs strongly across all four categories.

D.4 TSS Decomposition Analysis

Table 12 reports the TSS decomposition for each configuration. TSS credits a model only when it answers both the forward and backward question correctly for the same video instance, so it is a strict measure of genuine bidirectional understanding.

The full model achieves the highest Both Correct rate while simultaneously minimizing both Fwd Only and Bwd Only errors. This confirms that its gains arise from genuine bilateral causal understanding rather than one-direction shortcuts. SFT+TS-GRPO without TVCD yields the lowest Both Correct rate among the non-trivial configurations, which reinforces the need for inference-time contrastive decoding.

D.5 Summary

Table 13 gives a compact comparison across configurations. No partial configuration achieves best-in-class performance across all criteria simultaneously. The full ReVID model is the only configuration that maximizes TSS, excels on the causally demanding categories, and delivers consistent Overall gains over the backbone.

Table 11: Single-component gains over the LLaDA-V backbone on RevVideo. ΔOverall and ΔTSS denote absolute improvements over the no-component baseline. F/B Gap is the absolute difference between Forward and Backward accuracy.

Config	Fwd (%)	Bwd (%)	F/B Gap (pp)	Overall (%)	ΔOverall	TSS	ΔTSS
Backbone	51.89	49.25	2.64	50.57	–	30.53	–
SFT only	50.71	51.14	0.43	50.93	+0.36	30.77	+0.24
TS-GRPO only	53.77	51.57	2.20	52.67	+2.10	30.41	-0.12
TVCD only	53.85	52.92	0.93	53.38	+2.81	31.84	+1.31
Full (ReVID)	**59.83**	**58.05**	1.78	**58.94**	**+8.37**	**39.03**	**+8.50**

Table 12: TSS decomposition analysis. "Both Correct" is the fraction of instances where the model answers both forward and backward questions correctly. "Fwd Only" and "Bwd Only" denote instances where only one direction is correct.

Config	Both Corr. (%)	Fwd Only (%)	Bwd Only (%)	Both Wrong (%)	TSS	Ovrl. (%)
Backbone	30.53	21.36	18.72	29.39	30.53	50.57
SFT only	30.77	19.94	20.37	28.92	30.77	50.93
TS-GRPO only	30.41	23.36	21.16	25.07	30.41	52.67
TVCD only	31.84	22.01	21.08	25.07	31.84	53.38
SFT+TS-GRPO	25.07	22.58	23.08	29.27	25.07	47.65
TS-GRPO+TVCD	33.26	20.59	21.83	24.32	33.26	54.49
SFT+TVCD	29.20	20.41	20.73	29.66	29.20	49.61
Full (ReVID)	**39.03**	**20.80**	**19.02**	**21.15**	**39.03**	**58.94**

Taken together, these ablations show that all three components are jointly necessary and mutually reinforcing.

E Reward Component Ablation

To further investigate the two reward terms within Temporal-Symmetric GRPO, we remove each reward in isolation while keeping all other components fixed. The results are reported under the same evaluation protocol as the main paper.

Setup. We compare the full ReVID model with a variant without the Temporal Reliance Reward and a variant without the Unbiased Symmetry Reward.

Analysis. Both reward terms contribute meaningfully to the final performance, and removing either one causes consistent degradation across all metrics.

Effect of removing r_t. Dropping the Temporal Reliance Reward causes the largest single drop in Overall accuracy and TSS. This indicates that r_t is the more critical component for temporal coherence, because it explicitly rewards responses that depend on correct chronology rather than static frame semantics. The degradation is especially pronounced on Behavior Order and Dynamic Evolving, which require fine-grained causal tracking across time.

Effect of removing r_{ub}. Removing the Unbiased Symmetry Reward also leads to a substantial drop in Overall and TSS. The pattern differs from removing r_t: r_{ub} contributes more to Dynamic Evolving and Operation Series, where symmetric bidirectional attribution matters more than simple sequential prediction. This is exactly the intended role of the reward.

Summary. The two rewards are complementary and jointly necessary. r_t provides the main temporal grounding signal, while r_{ub} targets the directional imbalance that defines the Reversal Curse. The full ReVID model is the only configuration that simultaneously maximizes overall accuracy and TSS.

Table 13: Holistic comparison of the full model against key partial configurations.

Config	Highest TSS	Min F/B Gap	Best Causal Cats.	Overall > Backbone	Verdict
SFT only	×	✓	∼	∼	Marginal
TS-GRPO only	×	∼	×	∼	Partial
TVCD only	×	✓	×	∼	Partial
SFT+TS-GRPO	×	∼	∼	×	Insufficient
TS-GRPO+TVCD	∼	∼	∼	✓	Near-complete
Full (ReVID)	✓	∼	✓	✓	**Complete**

F TVCD Hyperparameters and Sensitivity

This section clarifies the TVCD hyperparameter setting used in the main results and reports a sensitivity study over contrast strength α and plausibility threshold β.

F.1 Hyperparameter Setting in the Main Results

For RevVideo results, we use a fixed TVCD configuration: contrast strength $\alpha = 1.0$ and plausibility threshold $\beta = 0.1$.

F.2 Sensitivity Analysis

To assess robustness, we conduct a grid search over $\alpha \in \{0.5, 1.0, 1.5\}$ and $\beta \in \{0.05, 0.10, 0.20\}$. Across all nine configurations, performance varies by less than 1.4 points in Overall accuracy and less than 1.2 in TSS, showing that TVCD is not sensitive to the precise parameter choice within the tested range. The best setting is also the one used in the main paper.

Table 14: Sensitivity of TVCD to contrast strength α and plausibility threshold β on RevVideo. Bold marks the configuration used in the main results.

α	β	Fwd (%)	Bwd (%)	Overall (%)	TSS
0.50	0.05	58.92	56.71	57.81	37.84
0.50	0.10	59.18	57.03	58.10	38.21
0.50	0.20	58.64	56.48	57.56	37.45
1.00	0.05	59.51	57.66	58.59	38.62
1.00	0.10	**59.83**	**58.05**	**58.94**	**39.03**
1.00	0.20	59.27	57.41	58.34	38.57
1.50	0.05	59.12	57.08	58.10	38.14
1.50	0.10	59.36	57.52	58.44	38.48
1.50	0.20	58.75	56.89	57.82	37.91

F.3 Implementation Note

TVCD modifies the logits using the contrastive view and then applies the adaptive plausibility constraint. Concretely, given logits $\mathbf{z}_{\text{orig}}$ and $\mathbf{z}_{\text{cd}}$ from the original and contrastive views, the decoding rule is

$$\mathbf{z} = (1 + \alpha)\mathbf{z}_{\text{orig}} - \alpha\mathbf{z}_{\text{cd}}, \tag{15}$$

followed by masking tokens that violate the plausibility threshold controlled by β. In all main experiments, this procedure uses $\alpha = 1.0$ and $\beta = 0.1$.

G Additional Clarifications and Design Rationale

G.1 Why the dVLM backbone matters

The core ReVID gain comes from combining a bidirectional dVLM backbone with bidirectional temporal supervision and inference-time contrastive decoding. To isolate the decoding rule, we also port TVCD to an AR backbone. The result improves over the plain AR model, but still falls short of ReVID, which shows that TVCD alone cannot substitute for a bidirectional backbone.

Table 15: TVCD applied to an AR backbone vs. ReVID. The contrastive rule ports to AR, but without bidirectional attention / iterative revision it does not match ReVID.

Method	Fwd↑	Bwd↑	Gap↓	Overall↑	TSS↑
LLaVA-NeXT-Video 7B	53.59	51.48	2.11	52.53	30.38
+ TVCD (AR)	55.70	54.01	**1.69**	54.85	32.91
ReVID (Ours, dLLM)	**59.83**	**58.05**	1.78	**58.94**	**39.03**

G.2 Why temporal shuffling is a valid counterfactual

The shuffled-video view used by TVCD is not meant to simulate literal reversal only. It is a stronger order-destroying counterfactual that suppresses chronology while preserving the same frames. Because the benchmark keeps clips with explicit temporal change, the model must rely on temporal structure rather than static appearance or language priors.

G.3 Efficiency and memory behavior

TVCD introduces no additional trainable parameters. Its cost is purely inference-time: the contrastive branch adds an extra forward pass per denoising step, which is why the latency overhead is stable across settings. The cache is maintained separately for the original and shuffled branches within one sample, and it is reset across samples to avoid leakage.