Video-Zero: Self-Evolution Video Understanding

Ruixu Zhang[1,§] **Deyi Ji**[2,‡] **Lanyun Zhu**[3,†] **Xuanyi Liu**[4]
Yuxin Meng[1] **Ruihang Chu**[1,†] **Yujiu Yang**[1,†]
[1]Tsinghua University [2]Tencent [3]Tongji University [4]Peking University

Abstract

Self-evolution offers a promising path for improving reasoning models without relying on intensive human annotation. However, extending this paradigm to video understanding remains underexplored and challenging: videos are long, dynamic, and redundant, while the evidence needed for reasoning is often sparse and temporally localized. Naively generating difficult question-answer pairs from full videos can therefore produce supervision that appears challenging but is weakly grounded, relying on static cues or language priors rather than temporal evidence. In this work, we argue that the key bottleneck of video self-evolution is not difficulty alone, but grounding. We propose **Video-Zero**, an annotation-free Questioner–Solver co-evolution framework that centers self-evolution on temporally localized evidence. The Questioner discovers informative evidence segments and generates evidence-grounded questions, while the Solver learns to answer and align its predictions with the supporting evidence. This closes an iterative loop of evidence discovery, grounded supervision, and evidence-aligned learning. Across 13 benchmarks spanning temporal grounding, long-video understanding, and video reasoning, Video-Zero consistently improves multiple video VLM backbones, demonstrating the effectiveness and transferability of evidence-centered self-evolution. Code and models are publicly available at https://github.com/Zzz99999/Video-Zero.

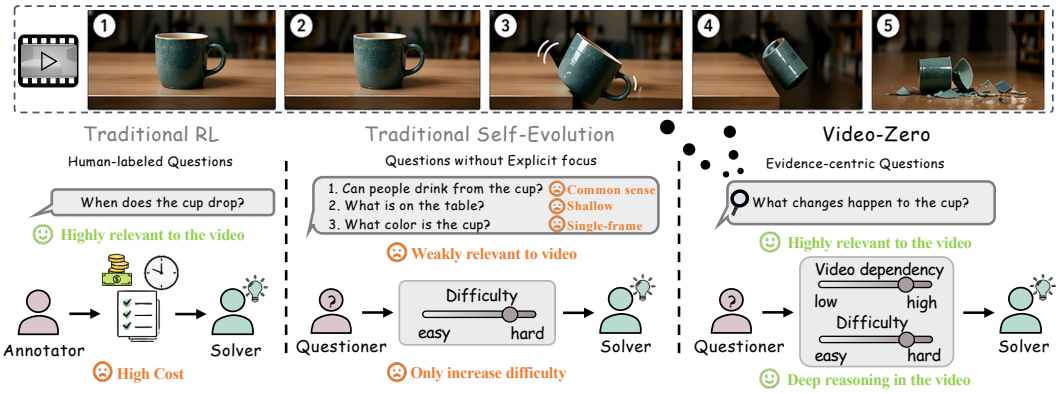
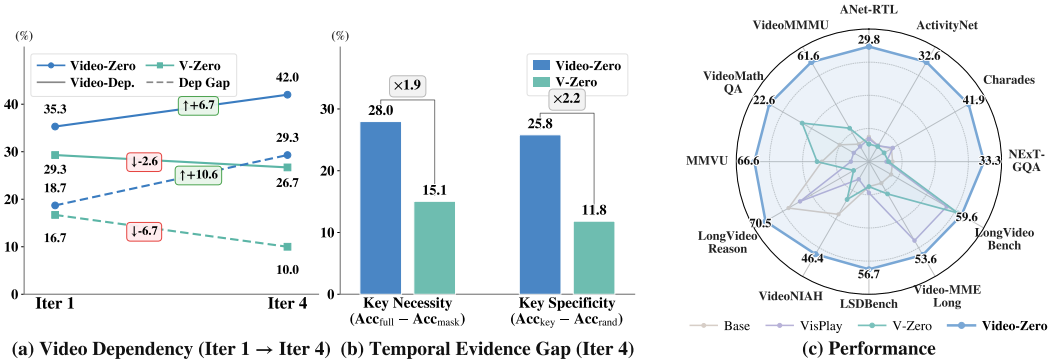


Figure 1: **Motivation and overview of Video-Zero.** Traditional RL relies on costly labels; prior self-evolution often increases difficulty without explicit evidence focus. Video-Zero grounds question generation in evidence segments and co-evolves toward video-dependent, challenging supervision.

[†]Corresponding authors.
[‡]Project leader.
[§]Work done during an internship at Tencent.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



(a) Video Dependency (Iter 1 → Iter 4) **(b) Temporal Evidence Gap (Iter 4)** **(c) Performance**

Figure 2: **Evidence-centered analysis and performance.** Video-Zero improves (a) video dependency, (b) key-span necessity and specificity in generated questions, and (c) performance across 13 benchmarks. Key Necessity (↑) measures whether the key span is indispensable, while Key Specificity (↑) measures whether it is more informative than a random span. See Appendix D.5 for details.

1 Introduction

Recent advances in Vision-Language Models (VLMs) suggest that strong reasoning capabilities can emerge not only from carefully human-annotated supervision, but also from iterative self-improvement [1–4]. In such self-evolution paradigms, models generate and learn from their own supervision signals within an iterative closed loop, typically instantiated by two co-evolving roles: a *Questioner* that constructs supervision from unlabeled data and a *Solver* that learns from the generated questions [5–8]. Despite promising progress in text and image domains, video understanding still lacks an effective annotation-free self-evolution framework. This gap is particularly critical, as high-quality video supervision data is both costly and scarce [9–11].

Achieving effective self-evolution for video understanding is non-trivial. Unlike text or images, videos are long, dynamic, and highly redundant, while task-relevant information required to construct supervision is often sparse and temporally localized [12]. This characteristic exposes two fundamental limitations when naively adapting current self-evolution approaches to the video domain. As shown in Figure 1, *(i)* questions are typically generated from the entire video without local focus, often resulting in queries that are weakly related to the actual video. Consequently, the trained model may rely only on language priors or static visual cues, rather than temporal dynamics necessary for effective video reasoning. This is further supported by the results in Figure 2, where the full-video generation method (V-Zero [7]) produces questions with weaker dependence on the video and lower sensitivity to key temporal evidence; *(ii)* existing self-evolution loops are primarily organized around question difficulty [7, 13], i.e., by optimizing the Questioner with a difficulty reward to avoid the generation of overly easy or hard samples. Yet, training effectiveness can be affected by multiple factors beyond difficulty, particularly the relevance between the generated question and the underlying video content. Therefore, we argue that effective self-evolution for video understanding requires a paradigm that goes beyond full-video generation and difficulty-driven optimization.

In this paper, we propose **Video-Zero**, a novel annotation-free self-evolution framework that enhances the video understanding capabilities of MLLMs directly from unlabeled videos. To address the above two limitations, Video-Zero features a fine-grained, temporally localized question-answer generation mechanism, along with a more comprehensive evidence-aware reward design beyond difficulty alone. Our key design lies in an *evidence-centric* manner, where temporal evidence (e.g., key events, transitions, and interactions) is explicitly inferred and leveraged for both supervision data generation and model optimization. Specifically, we develop a new question–answer generation mechanism conditioned on temporally localized evidence, encouraging the production of higher-quality supervision for more effective training, as evidenced by Figure 2(a)(b). In addition, we introduce a comprehensive reward design that incorporates multiple factors such as question–video relevance, improving optimization for both data generation and video understanding. By incorporating these integrated designs, the original video MLLM is improved through a co-evolution process: an enhanced Questioner produces more informative supervision, while a stronger Solver provides more reliable feedback for subsequent evidence selection and question construction.

We validate Video-Zero across 13 benchmarks using diverse model families, including powerful Qwen3-VL family, MiMo-VL, and InternVL3.5. Comprehensive experimental results demonstrate

that Video-Zero delivers significant and consistent performance gains across temporal grounding, long-video understanding, and video reasoning—all achieved without human annotation (Figure 2(c)). Our main contributions are as follows:

• We identify a fundamental bottleneck in video self-evolution, showing that useful supervision should be anchored to temporally localized evidence rather than driven by question difficulty alone.

• We propose Video-Zero, an *evidence-centered* co-evolution framework that unifies evidence discovery, grounded question construction, and evidence-aligned learning within a Questioner–Solver loop, requiring no human-written questions, answers, or temporal annotations.

• Extensive evaluations across multiple model architectures and 13 benchmarks deliver consistent and considerable improvements on diverse tasks, demonstrating the stability and scalability.

2 Related Work

Self-evolution in LLMs and VLMs. Self-evolution reduces reliance on human annotations by letting models generate, verify, filter, and learn from their own supervision signals [1, 2, 4, 14]. Recent reasoning-based and multimodal frameworks extend this paradigm through co-evolution, task proposal, verification, and supervision construction from unlabeled or synthetic visual inputs [5–8, 13, 15–19]. However, they remain largely task-centric, emphasizing difficulty while leaving temporal evidence grounding under-specified. Video-Zero addresses this gap by making temporally localized evidence the organizing unit for question generation, solver feedback, and iterative learning.

Video reasoning and temporal grounding. Recent video reasoning and post-training methods show that reinforcement learning can improve video MLLMs on video QA, temporal perception, and long-video reasoning [9–11, 20, 21]. A key trend is to encourage reasoning over relevant temporal moments through temporal rewards, evidence-aware reasoning traces, or temporally grounded supervision [9, 10, 22]. However, these pipelines typically rely on externally constructed supervision, such as curated reasoning data, cold-start traces, task-specific rewards, expert demonstrations, or temporal annotations, which are costly to scale for fine-grained video grounding [9, 11, 21, 22]. In contrast, Video-Zero generates supervision directly from unlabeled videos and organizes the learning loop around temporally informative evidence. We provide a more detailed discussion in Appendix B.

3 Method

3.1 Overview

We propose Video-Zero, an evidence-centered co-evolution framework for video self-evolution. Starting from unlabeled videos, Video-Zero constructs supervision signals using model-generated evidence, questions, and answers, and iteratively improves the model without human annotations. As illustrated in Figure 3, given an input video v with duration T, the framework maintains a Questioner policy π_Q and a Solver policy π_S, both initialized from the same pretrained backbone. For each video, the Questioner first identifies one or more informative temporal evidence spans $e = \{[t_k^s, t_k^e]\}_{k=1}^K$, where $0 \le t_k^s < t_k^e \le T$, and then constructs an evidence-grounded supervision unit in the form of a question-answer pair (q, a^Q) conditioned on (v, e). Given (v, q), the Solver predicts both an answer and a temporal span that localizes the evidence supporting its reasoning.

The key novelty of our Video-Zero lies in its evidence-centric design philosophy, reflected in three key components. *(i)* The Questioner constructs questions from localized temporal evidence rather than unconstrained video-level impressions. *(ii)* The Questioner is optimized with a video-aware utility that jointly considers learnability, video dependency, and evidence quality, since difficulty alone is insufficient in the video domain. *(iii)* To prevent answer shortcuts, the Solver is trained with pseudo answer labels and temporal targets derived from rollout consensus, along with a temporal alignment reward. These designs form a closed co-evolution loop where the Questioner provides better supervision and the Solver provides more reliable feedback for question generation.

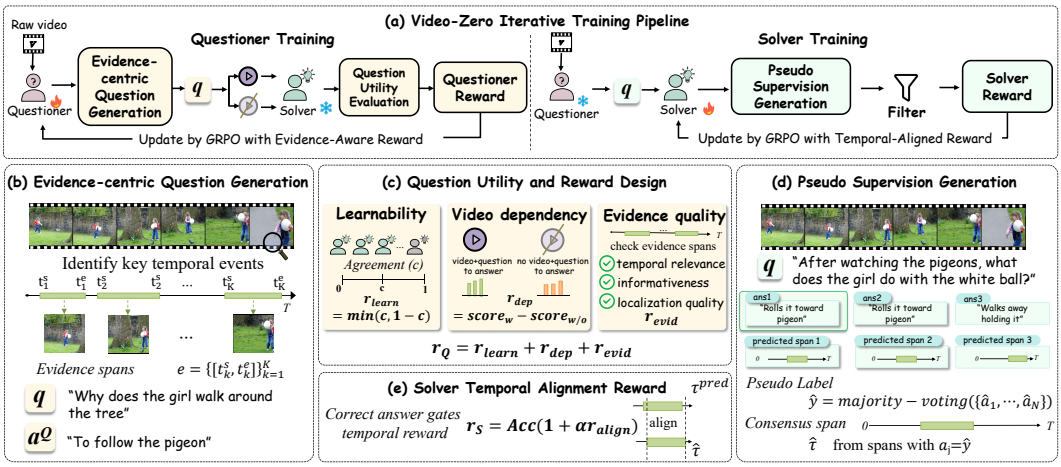


Figure 3: **Overview of Video-Zero.** (a) Video-Zero organizes video self-evolution around temporally localized evidence. A Questioner (b) discovers key evidence spans and generates evidence-grounded questions, which (c) are scored by learnability, video dependency, and evidence quality. A Solver then (d) learns from rollout pseudo supervision and (e) is optimized with a temporal alignment reward for evidence-aligned answer prediction, enabling closed-loop Questioner–Solver co-evolution.

3.2 Questioner: Evidence-Grounded Supervision Generation

3.2.1 Evidence-Centric Generation Policy

As illustrated in Figure 3(b), the Questioner first identifies temporally informative evidence and then constructs a question conditioned on that evidence. This turns question generation from an unconstrained process over raw videos into an evidence-guided process over localized temporal content. Formally, the Questioner implements this evidence-centric policy through the following factorization:

$$\pi_Q(e, q, a^Q \mid v) = \pi_Q(e \mid v)\pi_Q(q, a^Q \mid v, e), \tag{1}$$

where the temporal evidence e is generated before the question q and answer a^Q are finalized. The Questioner outputs structured samples containing temporal spans, evidence descriptions, question text, and answer candidates, making evidence identification explicit before question construction.

3.2.2 Question Utility and Reward Design

The evidence-centric policy only specifies how questions are generated, but does not determine which questions are useful for self-evolution. We thus define a question utility criterion to guide Questioner optimization. Prior self-evolution methods often rely on difficulty for this criterion; yet, in the video domain, a difficult question may still be answerable using language priors, answer patterns, or weak static cues, without requiring genuine temporal reasoning or reliance on video content. To address this, we propose a more comprehensive utility formulation with three dimensions (Fig. 3(c)): learnability, video dependency, and evidence quality, ensuring that a useful question is *(i)* learnable for the current Solver, *(ii)* dependent on video content, and *(iii)* grounded in high-quality temporal evidence.

Learnability. We estimate learnability using the Solver's rollout consistency under video input. For each candidate question, we query the current Solver m times with access to the video, obtaining sampled answers $\{a_1, \ldots, a_m\}$. Next, we compute the majority-vote answer $\hat{a}$ by counting occurrences of each answer in $\mathcal{A}$ among the sampled responses, define the consistency score $c \in [0, 1]$ as the fraction of responses matching $\hat{a}$, and then compute the learnability reward r_{learn} based on c. This process is formulated as follows:

$$\hat{a} = \arg\max_{a \in \mathcal{A}} \sum_{j=1}^{m} \mathbf{1}[a_j = a], \qquad c = \frac{1}{m} \sum_{j=1}^{m} \mathbf{1}[a_j = \hat{a}], \qquad r_{\text{learn}} = \min(c, 1 - c), \tag{2}$$

where $\mathbf{1}[\cdot]$ denotes the indicator function. Consistency serves as a proxy for whether a question lies near the current Solver's learnable boundary. Since r_{learn} is maximized at intermediate consistency, it

favors questions that are challenging but still learnable. In practice, we further downweight overly easy questions with a small penalty derived from r_{learn}.

Video dependency. Learnability assesses whether a question falls within an appropriate difficulty range for Solver's effective learning; however, it cannot distinguish genuinely video-dependent questions from those that can be solved just using non-video shortcuts. We therefore introduce a video-dependency reward that measures reliance on visual-temporal information by comparing the Solver's agreement with the Questioner-provided answer a^Q under with-video and without-video conditions. Specifically, we sample m Solver responses $\{a_j^{\text{with}}\}_{j=1}^m$ with video access and m responses $\{a_j^{\text{without}}\}_{j=1}^m$ without video access, and define:

$$\Delta_{\text{video}} = \frac{1}{m}\sum_{j=1}^{m} \mathbf{1}[a_j^{\text{with}} = a^Q] - \frac{1}{m}\sum_{j=1}^{m} \mathbf{1}[a_j^{\text{without}} = a^Q], \qquad r_{\text{dep}} = \max(\Delta_{\text{video}}, 0). \tag{3}$$

A larger positive Δ_{video} indicates that video input is more essential for solving the question, whereas a small or negative value suggests reliance on non-video shortcuts such as language priors, answer-option biases, static cues, or generic scene knowledge.

Evidence quality. Learnability and video dependency characterize question utility for the current Solver, but do not directly ensure high-quality evidence. A question may be learnable and video-dependent while relying on evidence that is vague, overly broad, or weakly tied to key events. We therefore introduce an evidence-quality reward computed from the structured Questioner output without external semantic labels: $r_{\text{evid}} = \text{clip}_{[-\beta,\beta]}(s_{\text{span}} + s_{\text{event}} + s_{\text{temp}} - \beta)$. The three components score the number of parsed temporal spans, the number of text segments in the evidence description, and the density of temporal expressions, respectively; detailed scoring rules are given in Appendix C.2.

Combining the above components, we define the final Questioner reward as

$$r_Q = r_{\text{format}} + \underbrace{(r_{\text{learn}} + r_{\text{dep}} + r_{\text{evid}})}_{U_Q}. \tag{4}$$

This reward r_Q consists of a format term r_{format} that enforces a valid output structure, and a utility term U_Q that aggregates signals indicative of high-value generated data. Note that prior to aggregation, each component is normalized to a comparable scale (implementation details are provided in the appendix). Together, these terms encourage the generated questions to be learnable for the current Solver, dependent on the video, and grounded in high-quality temporal evidence, thereby encouraging the Questioner to generate higher-quality data.

3.3 Solver Learning with Evidence-Aligned Supervision

The Questioner is executed and optimized based on temporal evidence. Similarly, we also construct the Solver in an evidence-centric manner to enhance its effectiveness. Specifically, the Solver is trained with evidence-aligned supervision by deriving pseudo answer labels and temporal targets from rollout consistency (Fig. 3(d)), and is further optimized using a temporal alignment reward (Fig. 3(e)).

Pseudo Supervision from Rollout Consistency. Given a Questioner-generated question q, we derive both pseudo answers and temporal spans from the Solver's rollout distribution conditioned on (v, q). Specifically, for each question, the current Solver samples n responses, where the j-th response includes an answer $f_j(v, q)$ along with a predicted temporal span $[t_s^{(j)}, t_e^{(j)}]$. We obtain a pseudo answer label via majority voting over the answer space $\mathcal{A}$ and use the vote ratio as a confidence measure. Temporal spans from rollouts that agree with the pseudo label are then aggregated to form a consensus temporal target:

$$\begin{aligned} \hat{y} &= \arg\max_{a\in\mathcal{A}} \sum_{j=1}^{n} \mathbf{1}[f_j(v,q) = a], \qquad s = \frac{1}{n}\max_{a\in\mathcal{A}} \sum_{j=1}^{n} \mathbf{1}[f_j(v,q) = a], \\ \hat{\tau} &= \left[\text{median}\Big(\{t_s^{(j)} : f_j(v,q) = \hat{y}\}\Big), \text{median}\Big(\{t_e^{(j)} : f_j(v,q) = \hat{y}\}\Big)\right]. \end{aligned} \tag{5}$$

Here, $\hat{y}$ is the majority-vote pseudo label, s denotes its support ratio, and $\hat{\tau}$ is the median start/end span among rollouts agreeing with $\hat{y}$. The confidence score s serves as both a reliability estimate and

a curriculum signal. We retain samples within an intermediate confidence range, filtering out those that are either trivially easy or excessively noisy.

Temporal Alignment Reward. We combine the obtained q, $\hat{y}$ and $\hat{\tau}$ as the pseudo supervision signal $\tilde{x}_S$. Given $\tilde{x}_S$, the Solver outputs a predicted answer $\hat{a}^p$ and temporal span $\tau^p = [t_s^p, t_e^p]$. We train it with a reward combining pseudo answer correctness and temporal alignment with $\hat{\tau} = [\hat{t}_s, \hat{t}_e]$:

$$r_S = \mathbf{1}[\hat{a}^p = \hat{y}]\,(1 + \alpha r_{\text{align}})\,, \qquad r_{\text{align}} = \text{IoU}(\tau^p, \hat{\tau}) \prod_{b \in \{s,e\}} \left[1 - \frac{|t_b^p - \hat{t}_b|}{T}\right]_+, \tag{6}$$

where $[x]_+ = \max(x, 0)$ and $\alpha = 0.5$ controls the temporal alignment contribution. The multiplicative form grants alignment reward only when the predicted answer matches the pseudo label, while the boundary term encourages start- and end-time consistency. In this way, the Solver learns not only to answer correctly, but also to localize the supporting temporal evidence.

3.4 Iterative Co-Evolution and Optimization

Video-Zero performs self-evolution through an alternating Questioner–Solver loop. At iteration t, the current Solver S_t serves as a fixed evaluator for Questioner optimization, encouraging supervision that is learnable, video-dependent, and grounded in high-quality temporal evidence. The optimized Questioner Q_{t+1} generates an evidence-grounded training set $\mathcal{D}_{t+1}$ from unlabeled videos. For each sample, Solver rollouts derive pseudo answer labels and temporal targets, and the Solver is trained on $\mathcal{D}_{t+1}$ with evidence-aligned supervision to obtain S_{t+1}. This forms the path $S_t \to Q_{t+1} \to \mathcal{D}_{t+1} \to S_{t+1}$, inducing an evidence-centered curriculum matched to the Solver's evolving capability. As the Solver improves, it provides more reliable feedback for evidence discovery and grounded question construction, enabling the Questioner to generate stronger supervision for the next update.

Both the Questioner and Solver are optimized with Group Relative Policy Optimization (GRPO) [23]. For each prompt x, we sample a group of outputs $\{o_i\}_{i=1}^G$, compute scalar rewards r_i, and form group-normalized advantages $A_i = (r_i - \mu_g)/(\sigma_g + \epsilon)$. The policy is then updated with the clipped GRPO objective and KL regularization. This shared RL procedure is applied to both roles, using the Questioner reward r_Q in Sec. 3.2.2 and the Solver reward r_S in Sec. 3.3.

4 Experiments

We evaluate Video-Zero along two axes: whether evidence-centered self-evolution improves temporal localization, and whether this transfers to broader video understanding and reasoning.

4.1 Experimental Setup

Models and baselines. Our main experiments use Qwen3-VL-4B-Instruct and Qwen3-VL-8B-Instruct [24] as the shared backbones for both the Questioner and Solver. We compare Video-Zero with the base model and two representative self-evolution baselines, VisPlay [13] and V-Zero [7], using matched budgets, iterative schedules, and evaluation protocols unless otherwise specified. We further test cross-architecture transfer on MiMo-SFT-7B [25] and InternVL3.5-4B [26].

Benchmarks. We evaluate Video-Zero on 13 benchmarks across temporal grounding, long-video understanding, and video reasoning. **Temporal grounding**: ANet-RTL [27], ActivityNet Captions [28], Charades-STA [29], and NExT-GQA [30]. **Long-video understanding**: LongVideoBench [31], MLVU [32], Video-MME-Long [33], LSDBench [34], and VideoNIAH [35]. **Video reasoning**: LongVideoReason [36], MMVU [37], VideoMathQA [38], and VideoMMMU [39].

Metrics and training. We report mIoU and Recall for ANet-RTL, ActivityNet, and Charades, GQA at IoU 0.3/0.5 for NExT-GQA, and official scores for long-video and reasoning benchmarks. For training, Video-Zero uses only 600 randomly sampled raw videos from Open-o3-Video [22], discarding all associated annotations. All methods are trained for 5 iterations on the same unlabeled videos under matched budgets and evaluated with the same protocols, with details in Appendix C.

Table 1: Main results on temporal grounding for Qwen3-VL-4B and Qwen3-VL-8B. The best and second-best results within each model block are highlighted in dark and light blue, respectively.

Methods	ANet-RTL				ActivityNet				Charades				NExT-GQA	
	mIoU	R@0.3	R@0.5	R@0.7	mIoU	R@0.3	R@0.5	R@0.7	mIoU	R@0.3	R@0.5	R@0.7	GQA@0.3	GQA@0.5
Qwen3-VL-4B-Instruct														
Base Model	16.86	22.70	15.72	8.73	22.86	30.94	19.04	11.29	37.52	58.90	36.12	16.23	23.09	15.40
VisPlay	16.58	23.14	13.97	8.73	22.89	30.98	19.12	11.31	37.59	58.82	36.59	16.42	23.14	15.33
V-Zero	15.85	20.96	13.97	8.30	22.81	30.91	18.92	11.31	37.10	58.33	35.24	15.62	22.82	15.22
Video-Zero (Iter 1)	20.86	28.82	20.09	8.73	24.23	32.97	20.22	12.04	38.75	60.59	36.91	17.45	31.66	20.80
Video-Zero (Iter 2)	23.81	33.19	22.27	11.35	27.30	37.48	23.16	13.66	40.85	64.19	39.81	18.60	36.59	22.80
Video-Zero (Iter 3)	27.75	39.30	28.82	13.97	30.29	41.82	26.15	15.50	41.66	64.54	41.40	19.60	38.14	23.25
Video-Zero (Iter 4)	29.75	44.98	28.82	13.54	32.12	44.76	27.67	16.24	41.85	65.11	40.81	20.03	38.29	23.00
Video-Zero (Iter 5)	29.34	42.79	27.95	13.54	32.59	45.35	28.01	16.38	41.48	64.62	39.92	19.95	37.76	22.82
Peak Δ	**+12.89**	**+22.28**	**+13.10**	**+5.24**	**+9.73**	**+14.41**	**+8.97**	**+5.09**	**+4.33**	**+6.21**	**+5.28**	**+3.80**	**+15.20**	**+7.85**
Qwen3-VL-8B-Instruct														
Base Model	21.95	30.57	21.83	10.04	26.56	36.59	22.63	12.85	36.70	59.11	33.87	13.49	35.93	21.83
VisPlay	21.83	32.31	20.09	10.04	26.74	36.94	22.99	12.93	36.65	59.03	33.92	13.47	35.87	22.19
V-Zero	22.77	32.31	20.52	9.17	26.69	36.79	23.05	13.00	36.43	58.63	33.66	13.41	35.73	22.11
Video-Zero (Iter 1)	24.78	34.50	24.45	12.23	26.75	36.87	23.08	12.99	36.32	58.58	32.93	12.82	37.98	23.55
Video-Zero (Iter 2)	23.90	35.37	23.14	12.23	27.79	38.49	23.94	13.42	36.37	58.84	32.77	12.72	39.04	24.31
Video-Zero (Iter 3)	24.48	37.99	21.83	11.79	28.99	40.17	25.20	14.15	36.84	59.22	33.60	13.12	39.29	24.24
Video-Zero (Iter 4)	26.19	39.30	24.45	12.23	29.87	41.77	26.23	14.66	37.34	60.16	35.54	13.58	39.60	24.26
Video-Zero (Iter 5)	23.74	34.50	21.40	9.61	29.42	41.00	25.83	14.39	37.07	60.03	34.41	12.93	39.96	25.10
Peak Δ	**+4.24**	**+8.73**	**+2.62**	**+2.19**	**+3.31**	**+5.18**	**+3.60**	**+1.81**	**+0.64**	**+1.05**	**+1.67**	**+0.09**	**+4.03**	**+3.27**

Table 2: Main results on long-video understanding and video reasoning benchmarks for Qwen3-VL-4B and Qwen3-VL-8B. The best and second-best results within each model block are highlighted in dark and light blue, respectively.

Methods	Long-video Understanding					Video Reasoning			
	LongVideoBench	MLVU	Video-MME-L	LSDBench	VideoNIAH	LongVideoReason	MMVU	VideoMathQA	VideoMMMU
Qwen3-VL-4B-Instruct									
Base Model	57.70	62.60	51.60	54.68	45.78	69.90	64.32	19.05	57.56
VisPlay	59.20	63.50	53.20	54.83	45.26	69.60	63.36	18.33	57.44
V-Zero	59.50	63.50	51.90	54.68	45.56	68.20	64.48	20.95	58.33
Video-Zero (Iter 4)	59.10	63.40	53.60	56.52	46.37	70.30	64.96	21.67	61.56
Video-Zero (Peak)	59.60	63.40	53.60	56.67	46.37	70.50	66.56	22.62	61.56
Qwen3-VL-8B-Instruct									
Base Model	59.80	64.40	55.70	57.59	45.70	68.60	68.32	23.81	63.78
VisPlay	60.50	64.90	54.90	57.44	45.70	68.10	67.04	21.43	62.67
V-Zero	60.10	64.20	56.00	56.90	46.07	69.20	68.80	22.14	63.22
Video-Zero (Iter 4)	60.40	64.60	56.60	56.75	46.37	70.70	68.00	25.48	64.22
Video-Zero (Peak)	60.60	64.70	57.70	58.13	46.37	71.00	68.32	26.19	65.33

4.2 Main Results

Temporal grounding. We first examine whether organizing video self-evolution around temporal evidence improves temporally grounded video understanding. As shown in Table 1, Video-Zero consistently improves both Qwen3-VL-4B and Qwen3-VL-8B across temporal localization benchmarks and grounded video QA, whereas prior self-evolution baselines such as VisPlay and V-Zero yield smaller or less consistent gains. This suggests that temporal grounding benefits not merely from additional self-generated video QA supervision, but more directly from supervision explicitly organized around the video evidence that supports each answer. The gains are especially pronounced on ANet-RTL and ActivityNet, where identifying the correct temporal segment is central to performance. It also substantially improves NExT-GQA, indicating that the model also learns to connect answers with their supporting temporal evidence rather than only predict better timestamps.

Video QA and reasoning. We next examine whether Video-Zero's temporal-evidence modeling transfers beyond explicit grounding. As shown in Table 2, Video-Zero brings broad improvements across long-video understanding and video reasoning benchmarks on both backbones. Since these tasks do not directly reward timestamps, the gains suggest that evidence-centric self-evolution improves the ability to more reliably identify, aggregate, and reason over temporally distributed video information. Video-Zero improves most long-video benchmarks, including LongVideoBench, Video-MME-Long, LSDBench, and VideoNIAH, while remaining competitive on MLVU. It also

yields clear gains on reasoning benchmarks such as VideoMathQA and VideoMMMU, where models must combine video content with higher-level reasoning rather than recognize events. Together, these results show that temporal evidence provides a transferable interface for video self-evolution.

4.3 Cross-Architecture Transfer

We further examine whether Video-Zero generalizes beyond the Qwen3-VL backbones used in our main experiments. This setting tests whether evidence-centered self-evolution is tied to a particular model family or can serve as a more general training mechanism. As shown in Table 3, we apply the same framework to MiMo-SFT-7B and InternVL3.5-4B without changing the pipeline. Both backbones also improve on most benchmarks, indicating that the gains are not specific to the Qwen-VL family. The transfer is especially clear on reasoning-intensive tasks: MiMo-SFT-7B improves across all video reasoning benchmarks and most long-video understanding benchmarks, while InternVL3.5-4B obtains large gains on MMVU, VideoMathQA, and VideoMMMU despite mixed long-video results. These results suggest that Video-Zero transfers across architectures rather than being a backbone-specific recipe. The consistent gains support our central design choice: organizing self-generated supervision around temporal evidence provides a model-agnostic signal for evidence identification and use over time.

Table 3: Cross-architecture generalization results.

Benchmark	MiMo-SFT-7B		InternVL3.5-4B	
	Base	Video-Zero	Base	Video-Zero
Long-video Understanding				
LongVideoBench	47.60	53.00	60.30	61.20
MLVU	40.40	43.70	62.40	62.00
Video-MME-L	33.10	35.90	52.40	53.70
LSDBench	51.92	55.83	60.35	57.52
VideoNIAH	46.67	46.81	38.81	36.89
Video Reasoning				
LongVideoReason	69.10	73.10	68.00	72.00
MMVU	69.60	70.08	54.08	58.88
VideoMathQA	29.76	32.62	12.62	23.81
VideoMMMU	67.33	68.33	44.22	53.11

5 Analysis

Ablation study. We ablate Questioner-side, Solver-side, and curriculum components in the evidence-centric co-evolution framework. Table 4 reports ablations on Qwen3-VL-4B, with all variants evaluated at Iteration 4. The full model achieves the best grounding, video QA, reasoning, and overall averages, indicating that the gains come from the complete design. A key factor is supervision generation: replacing the evidence-centric Questioner with a description-only variant causes a large drop, especially on temporal grounding. This shows that generic video descriptions are insufficient; the Questioner must identify where the answer is supported so that generated questions remain tied to temporal evidence. On the Questioner side, removing video dependency, evidence quality, or learnability control hurts performance, showing that useful supervision should require visual evidence, be grounded in informative temporal spans, and remain learnable for the Solver. On the Solver side, removing the temporal reward causes the largest overall degradation, suggesting that answer correctness alone is insufficient and the Solver must align predictions with supporting evidence. Finally, freezing the Questioner underperforms the full model, confirming that the gains arise from iterative Questioner–Solver co-evolution rather than a fixed synthetic dataset. Together, these results support the central design of Video-Zero: evidence should guide both supervision construction and the learning signal used to absorb it.

Table 4: Ablation study on Qwen3-VL-4B.

Method	Grounding Avg.	Video QA & Reasoning Avg.	Overall Avg.
Video-Zero (full)	**33.02**	**55.28**	**48.43**
Questioner-side ablations			
Description-only Questioner	23.43	54.68	45.06
w/o Video Dependency	30.36	54.76	47.25
w/o Evidence Quality	31.54	54.76	47.62
w/o Learnability	29.29	54.90	47.02
Solver-side ablations			
w/o Temporal Reward	22.86	53.70	44.21
Curriculum / loop ablations			
Frozen Questioner	31.17	54.49	47.31

Iterative co-evolution dynamics. A key question is whether the gains of Video-Zero accumulate through iterative co-evolution or mainly come from a single round of self-generated supervision. Since temporal grounding is the capability most directly targeted by evidence-centric self-evolution, we track the average progressive grounding improvement over the base model across five iterations.

Table 5: **Questioner analysis across iterations.** Later Questioners generate more video-dependent and challenging questions. (a) Video dependency is reflected by the growing gap between with-video and without-video Solver inputs. (b) Solver-checkpoint evaluation further shows that later question sets tend to best match Solvers from similar iterations, while pseudo-label accuracy decreases.

(a) Video Dependency

Q Set	w/ Vid	w/o Vid	Δ_{video}
QIter1	56.50	49.00	+7.50
QIter2	50.00	40.50	+9.50
QIter3	52.50	42.50	+10.00
QIter4	54.00	37.50	+16.50
QIter5	49.50	30.50	**+19.00**

(b) Difficulty and Pseudo-label Reliability

Q Set	Base	S(1)	S(2)	S(3)	S(4)	S(5)	PL Acc.
QIter1	59.00	**66.00**	61.50	60.00	61.00	60.50	**64.50**
QIter2	46.50	52.00	**54.50**	50.00	50.50	54.00	57.50
QIter3	41.50	45.50	45.50	**49.50**	45.50	47.50	50.00
QIter4	37.50	34.50	32.50	33.50	**44.00**	39.00	40.50
QIter5	33.00	35.00	33.00	30.50	33.50	**36.00**	37.50

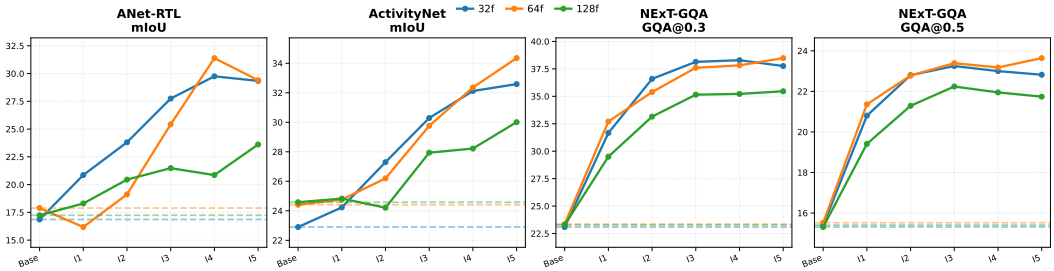


Figure 5: **Frame-budget robustness.** Video-Zero improves temporal grounding under 32/64/128 max-frame budgets at 2 FPS. Dashed lines denote base performance, and I denotes iteration.

As shown in Figure 4, both backbones continue to improve beyond the first iteration, indicating that Video-Zero is not merely a one-shot data generation procedure. The 4B model obtains larger iterative gains, while the 8B model improves more moderately, consistent with its stronger initialization. Although the curves are not strictly monotonic and begin to saturate in later rounds, their overall upward trend supports iterative Questioner–Solver co-evolution as a driver of temporal evidence localization. We analyze this behavior next by examining how the evolving Questioner changes the learnability, video dependency, and reliability of the generated supervision.

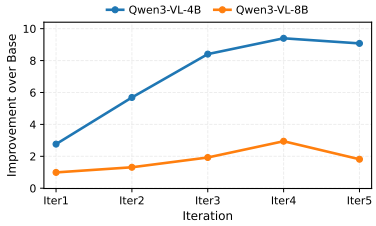


Figure 4: Iteration improvement in temporal grounding.

What does the Questioner learn? We further analyze how the Questioner evolves over iterations. All results are computed on 200 randomly sampled examples, with pseudo-label reliability measured against Qwen3-VL-235B-A22B [24]. As shown in Table 5, later Questioners generate supervision that is more video-dependent and challenging. The video dependency gap between with-video and without-video settings increases from +7.5 on QIter1 to +19.0 on QIter5, indicating reduced reliance on language priors or option patterns. Meanwhile, base-Solver accuracy on the generated question sets drops from 59.0 to 33.0, showing that the Questioner progressively moves beyond easy supervision. The best-performing Solver shifts with the Questioner iteration, forming a near-diagonal pattern from QIter1/S(1) to QIter5/S(5), suggesting that the Questioner generates questions near the contemporary Solver's learning boundary. However, this introduces a trade-off: as questions become harder and more video-dependent, pseudo-label accuracy drops from 64.5 on QIter1 to 37.5 on QIter5. This helps explain the saturation in Figure 4: early iterations benefit from more informative supervision, while later iterations face growing pseudo-label noise that partially offsets further gains.

Frame-Budget Robustness. Figure 5 examines whether the temporal grounding gains of Video-Zero depend on the visual context budget. We vary max sampled frames among 32, 64, and 128 while keeping the sampling rate fixed at 2 FPS. Across ANet-RTL, ActivityNet, and NExT-GQA, Video-Zero consistently improves over the corresponding base model under all three budgets, showing that the gains are not tied to a single frame budget. The improvements are especially clear under 32f and 64f, while 128f also remains generally above its base performance. Together with the additional robustness, scaling, seed, and qualitative analyses in Appendix D.6, these results suggest that evidence-centered self-evolution remains stable across different training and evaluation conditions.

Effect of training video duration. To examine the effect of training-video composition, we construct **LongMix** by replacing 200 short videos with 200 unlabeled videos of 5–20 minutes, keeping 600 videos in total. Both training variants and all evaluations use a maximum of 64 frames, with other training settings unchanged.

Table 6: **Effect of training-video composition under a fixed 64-frame budget.**

Setting	Long-Video Avg.	Reasoning Avg.	Charades mIoU	NExT-GQA @0.3	NExT-GQA @0.5
Base	57.73	52.54	38.77	23.36	15.52
Video-Zero (600 short)	58.34	54.00	**40.81**	**37.82**	23.18
Video-Zero (LongMix)	**58.75**	**55.33**	40.79	36.63	**23.54**

Table 6 shows that LongMix increases the mean gain over the base model from +0.61 to +1.02 across four long-video benchmarks and from +1.46 to +2.79 across four reasoning benchmarks. Grounding performance remains above the base model, with mixed changes relative to short-only training. These results suggest that incorporating longer unlabeled videos broadens the benefits of self-evolution beyond temporal grounding. Detailed settings and results are provided in Appendix D.6.2.

6 Conclusion

In this work, we studied annotation-free self-evolution for video understanding and identified a key bottleneck: useful supervision should be grounded in temporally localized evidence rather than driven by question difficulty alone. To address this challenge, we proposed **Video-Zero**, an evidence-centered Questioner–Solver co-evolution framework that improves video MLLMs directly from unlabeled videos. Video-Zero discovers more informative evidence segments, constructs grounded question-answer supervision, and trains the Solver to answer questions while aligning its predictions with the supporting temporal evidence. This iterative loop enables more reliable and transferable video self-evolution. Across 13 benchmarks and multiple video MLLM backbones, Video-Zero consistently improves temporal grounding, long-video understanding, and video reasoning, validating the effectiveness of organizing video self-evolution around temporal evidence.

Acknowledgments and Disclosure of Funding

This work was partly supported by the National Key Research and Development Program of China (No.2024YFB2808903), the Shenzhen Science and Technology Program (ZDCY20250901103533010).

References

[1] Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models. *CoRR*, abs/2401.10020, 2024.

[2] Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. Star: Bootstrapping reasoning with reasoning. In *NeurIPS*, 2022.

[3] Zhengwei Tao, Ting-En Lin, Xiancai Chen, Hangyu Li, Yuchuan Wu, Yongbin Li, Zhi Jin, Fei Huang, Dacheng Tao, and Jingren Zhou. A survey on self-evolution of large language models. *CoRR*, abs/2404.14387, 2024.

[4] Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning converts weak language models to strong language models. In *ICML*, Proceedings of Machine Learning Research, pages 6621–6642. PMLR / OpenReview.net, 2024.

[5] Andrew Zhao, Yiran Wu, Yang Yue, Tong Wu, Quentin Xu, Yang Yue, Matthieu Lin, Shenzhi Wang, Qingyun Wu, Zilong Zheng, and Gao Huang. Absolute zero: Reinforced self-play reasoning with zero data. *CoRR*, abs/2505.03335, 2025.

[6] Chengsong Huang, Wenhao Yu, Xiaoyang Wang, Hongming Zhang, Zongxia Li, Ruosen Li, Jiaxin Huang, Haitao Mi, and Dong Yu. R-zero: Self-evolving reasoning LLM from zero data. *CoRR*, abs/2508.05004, 2025.

[7] Han Wang, Yi Yang, Jingyuan Hu, Minfeng Zhu, and Wei Chen. V-zero: Self-improving multimodal reasoning with zero annotation. *CoRR*, abs/2601.10094, 2026.

[8] Zongxia Li, Hongyang Du, Chengsong Huang, Xiyang Wu, Lantao Yu, Yicheng He, Jing Xie, Xiaomin Wu, Zhichao Liu, Jiarui Zhang, and Fuxiao Liu. Mm-zero: Self-evolving multi-model vision language models from zero data. *CoRR*, abs/2603.09206, 2026.

[9] Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *CoRR*, abs/2503.21776, 2025.

[10] Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *CoRR*, abs/2504.06958, 2025.

[11] Kaituo Feng, Manyuan Zhang, Hongyu Li, Kaixuan Fan, Shuang Chen, Yilei Jiang, Dian Zheng, Peiwen Sun, Yiyuan Zhang, Haoze Sun, Yan Feng, Peng Pei, Xunliang Cai, and Xiangyu Yue. Onethinker: All-in-one reasoning model for image and video. *CoRR*, abs/2512.03043, 2025.

[12] Haibo Wang, Zhiyang Xu, Yu Cheng, Shizhe Diao, Yufan Zhou, Yixin Cao, Qifan Wang, Weifeng Ge, and Lifu Huang. Grounded-videollm: Sharpening fine-grained temporal grounding in video large language models. In *EMNLP (Findings)*, pages 959–975. Association for Computational Linguistics, 2025.

[13] Yicheng He, Chengsong Huang, Zongxia Li, Jiaxin Huang, and Yonghui Yang. Visplay: Self-evolving vision-language models from images. *CoRR*, abs/2511.15661, 2025.

[14] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In *ACL (1)*, pages 13484–13508. Association for Computational Linguistics, 2023.

[15] Bo Liu, Chuanyang Jin, Seungone Kim, Weizhe Yuan, Wenting Zhao, Ilia Kulikov, Xian Li, Sainbayar Sukhbaatar, Jack Lanchantin, and Jason Weston. SPICE: self-play in corpus environments improves reasoning. *CoRR*, abs/2510.24684, 2025.

[16] Ziyi Yang, Weizhou Shen, Ruijun Chen, Chenliang Li, Fanqi Wan, Ming Yan, Xiaojun Quan, and Fei Huang. SPELL: self-play reinforcement learning for evolving long-context language models. *CoRR*, abs/2509.23863, 2025.

[17] Shengda Fan, Xuyan Ye, and Yankai Lin. DARC: decoupled asymmetric reasoning curriculum for LLM evolution. *CoRR*, abs/2601.13761, 2026.

[18] Qinsi Wang, Bo Liu, Tianyi Zhou, Jing Shi, Yueqian Lin, Yiran Chen, Hai Helen Li, Kun Wan, and Wentian Zhao. Vision-zero: Scalable VLM self-improvement via strategic gamified self-play. *CoRR*, abs/2509.25541, 2025.

[19] Jinghan He, Junfeng Fang, Feng Xiong, Zijun Yao, Fei Shen, Haiyun Guo, Jinqiao Wang, and Tat-Seng Chua. Active zero: Self-evolving vision-language models through active environment exploration. *CoRR*, abs/2602.11241, 2026.

[20] 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. *arXiv preprint arXiv:2503.13377*, 2025.

[21] Yang Ding, Yizhen Zhang, Xin Lai, Ruihang Chu, and Yujiu Yang. Videozoomer: Reinforcement-learned temporal focusing for long video reasoning. *CoRR*, abs/2512.22315, 2025.

[22] Jiahao Meng, Xiangtai Li, Haochen Wang, Yue Tan, Tao Zhang, Lingdong Kong, Yunhai Tong, Anran Wang, Zhiyang Teng, Yujing Wang, and Zhuochen Wang. Open-o3 video: Grounded video reasoning with explicit spatio-temporal evidence. *CoRR*, abs/2510.20579, 2025.

[23] DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948, 2025.

[24] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[25] Zihao Yue, Zhenru Lin, Yifan Song, Weikun Wang, Shuhuai Ren, Shuhao Gu, Shicheng Li, Peidian Li, Liang Zhao, Lei Li, Kainan Bao, Hao Tian, Hailin Zhang, Xiao-Gang Wang, Dawei Zhu, Cici, Chenhong He, Bowen Ye, Bowen Shen, Zihan Zhang, Zihan Jiang, Zhixian Zheng, Zhichao Song, Zhenbo Luo, Yue Yu, Yudong Wang, Yuanyuan Tian, Yu Tu, Yihan Yan, Yi Huang, Xu Wang, Xinzhe Xu, Xingchen Song, Xing Zhang, Xing Yong, Xin Zhang, Xiangwei Deng, Wenyu Yang, Wenhan Ma, Weiwei Lv, Weiji Zhuang, Wei Liu, Sirui Deng, Shuo Liu, Shimao Chen, Shihua Yu, Shaohui Liu, Shande Wang, Rui Ma, Qiantong Wang, Peng Wang, Nuo Chen, Menghang Zhu, Kangyang Zhou, Kang Zhou, Kai Fang, Jun Shi, Jinhao Dong, Jiebao Xiao, Jiaming Xu, Huaqiu Liu, Hongshen Xu, Heng Qu, Haochen Zhao, Hanglong Lv, Guoan Wang, Duo Zhang, Dong Zhang, Di Zhang, Chong Ma, Chang Liu, Can Cai, and Bingquan Xia. Mimo-vl technical report. *CoRR*, abs/2506.03569, 2025.

[26] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

[27] De-An Huang, Shijia Liao, Subhashree Radhakrishnan, Hongxu Yin, Pavlo Molchanov, Zhiding Yu, and Jan Kautz. Lita: Language instructed temporal-localization assistant. In *European Conference on Computer Vision*, pages 202–218. Springer, 2024.

[28] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715, 2017.

[29] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275, 2017.

[30] Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. Can i trust your answer? visually grounded video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13204–13214, 2024.

[31] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024.

[32] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: Benchmarking multi-task long video understanding. *arXiv preprint arXiv:2406.04264*, 2024.

[33] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24108–24118, 2025.

[34] Tianyuan Qu, Longxiang Tang, Bohao Peng, Senqiao Yang, Bei Yu, and Jiaya Jia. Does your vision-language model get lost in the long video sampling dilemma? *arXiv preprint arXiv:2503.12496*, 2025.

[35] Zijia Zhao, Haoyu Lu, Yuqi Huo, Yifan Du, Tongtian Yue, Longteng Guo, Bingning Wang, Weipeng Chen, and Jing Liu. Needle in a video haystack: A scalable synthetic evaluator for video mllms. *arXiv preprint arXiv:2406.09367*, 2024.

[36] Yukang Chen, Wei Huang, Baifeng Shi, Qinghao Hu, Hanrong Ye, Ligeng Zhu, Zhijian Liu, Pavlo Molchanov, Jan Kautz, Xiaojuan Qi, Sifei Liu, Hongxu Yin, Yao Lu, and Song Han. Scaling RL to long videos. *CoRR*, abs/2507.07966, 2025.

[37] Yilun Zhao, Haowei Zhang, Lujing Xie, Tongyan Hu, Guo Gan, Yitao Long, Zhiyuan Hu, Weiyuan Chen, Chuhan Li, Zhijian Xu, et al. Mmvu: Measuring expert-level multi-discipline video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8475–8489, 2025.

[38] Hanoona Rasheed, Abdelrahman Shaker, Anqi Tang, Muhammad Maaz, Ming-Hsuan Yang, Salman Khan, and Fahad Shahbaz Khan. Videomathqa: Benchmarking mathematical reasoning via multimodal understanding in videos. *arXiv preprint arXiv:2506.05349*, 2025.

[39] Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. *arXiv preprint arXiv:2501.13826*, 2025.

[40] Yuran Wang, Zekun Wang, Bohan Zeng, Ruixu Zhang, Wenxuan Liu, Liu Yang, Yifan Dai, Yang Shi, Bozhou Li, Chengzhuo Tong, Daili Hua, Yuanxing Zhang, and Wentao Zhang. Flux-opd: On-policy distillation with evolving contexts. *CoRR*, abs/2607.28022, 2026.

[41] Cheng Chen, Jiacheng Wei, Tianrun Chen, Chi Zhang, Xiaofeng Yang, Shangzhan Zhang, Bingchen Yang, Chuan-Sheng Foo, Guosheng Lin, Qixing Huang, and Fayao Liu. Cadcrafter: Generating computer-aided design models from unconstrained images. In *CVPR*, pages 11073–11082. Computer Vision Foundation / IEEE, 2025.

[42] Wenhao Shen, Wanqi Yin, Xiaofeng Yang, Cheng Chen, Chaoyue Song, Zhongang Cai, Lei Yang, Hao Wang, and Guosheng Lin. ADHMR: aligning diffusion-based human mesh recovery via direct preference optimization. In *ICML*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025.

[43] Omkar Thawakar, Shravan Venkatraman, Ritesh Thawkar, Abdelrahman Shaker, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, and Fahad Khan. Evolmm: Self-evolving large multimodal models with continuous rewards. *arXiv preprint arXiv:2511.16672*, 2025.

[44] Xiuwei Chen, Wentao Hu, Hanhui Li, Jun Zhou, Zisheng Chen, Meng Cao, Yihan Zeng, Kui Zhang, Yu-Jie Yuan, Jianhua Han, Hang Xu, and Xiaodan Liang. C2-evo: Co-evolving multimodal data and model for self-improving reasoning. *CoRR*, abs/2507.16518, 2025.

[45] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025.

[46] Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyu Zeng, Chenting Wang, Changlian Ma, Haian Huang, Jianfei Gao, et al. Internvideo2. 5: Empowering video mllms with long and rich context modeling. *arXiv preprint arXiv:2501.12386*, 2025.

[47] Canhui Tang, Zifan Han, Hongbo Sun, Sanping Zhou, Xuchong Zhang, Xin Wei, Ye Yuan, Jinglin Xu, and Hao Sun. Tspo: Temporal sampling policy optimization for long-form video language understanding. *arXiv preprint arXiv:2508.04369*, 2025.

[48] Yongxin Guo, Jingyu Liu, Mingda Li, Qingbin Liu, Xi Chen, and Xiaoying Tang. TRACE: Temporal Grounding Video LLM via Causal Event Modeling. *arXiv preprint arXiv:2410.05643*, 2024.

[49] Shijian Wang, Jiarui Jin, Xingjian Wang, Linxin Song, Runhao Fu, Hecheng Wang, Zongyuan Ge, Yuan Lu, and Xuelian Cheng. Video-thinker: Sparking "thinking with videos" via reinforcement learning. *CoRR*, abs/2510.23473, 2025.

[50] Fuwen Luo, Shengfeng Lou, Chi Chen, Ziyue Wang, Chenliang Li, Weizhou Shen, Jiyue Guo, Peng Li, Ming Yan, Ji Zhang, Fei Huang, and Yang Liu. MUSEG: reinforcing video temporal understanding via timestamp-aware multi-segment grounding. *CoRR*, abs/2505.20715, 2025.

[51] Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. Vtimellm: Empower llm to grasp video moments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14271–14280, 2024.

[52] Xiangyu Zeng, Kunchang Li, Chenting Wang, Xinhao Li, Tianxiang Jiang, Ziang Yan, Songze Li, Yansong Shi, Zhengrong Yue, Yi Wang, et al. Timesuite: Improving mllms for long video understanding via grounded tuning. *arXiv preprint arXiv:2410.19702*, 2024.

[53] Ruixu Zhang, Yuran Wang, Xinyi Hu, Chaoyu Mai, Wenxuan Liu, Danni Xu, Xian Zhong, and Zheng Wang. Beyond the individual: Introducing group intention forecasting with SHOT dataset. In *Proceedings of the ACM International Conference on Multimedia*, pages 13002–13008. ACM, 2025.

[54] Xinyi Hu, Yuran Wang, Ruixu Zhang, Yue Li, Wenxuan Liu, and Zheng Wang. SPAN: continuous modeling of suspicion progression for temporal intention localization. In *ACM Multimedia*, pages 315–323. ACM, 2025.

[55] Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11198–11201, 2024.

A Appendix Overview

This appendix provides more supplementary material for Video-Zero. Appendix B expands the discussion of related work. Appendix C provides more implementation, training, and evaluation details. Appendix D reports more results, ablations, and qualitative analyses.

B Extended Related Work

Self-evolution in LLMs and VLMs. Models can improve through self-generated supervision, feedback-driven distillation [40], and automatically constructed preferences [41, 42], reducing reliance on manual annotation. Early work in language models explores self-generated instructions, rationales, rewards, and iterative policy improvement [1, 2, 4, 14]. Recent reasoning-oriented and multimodal frameworks extend this paradigm through co-evolution, task proposal, verification, and supervision construction from unlabeled or synthetic visual inputs [5–8, 13, 15–19, 43, 44]. These methods establish a general closed-loop recipe in which models propose tasks, evaluate their utility, and learn from retained examples. However, existing self-evolution methods are still largely task-centric, with reward criteria focusing on difficulty. While such criteria are effective for text and image reasoning, they do not explicitly ensure that a generated video question depends on temporally localized visual evidence. As a result, a question may appear plausible or difficult while still being answerable from language priors, static cues, or generic scene knowledge. Video-Zero addresses this limitation by making temporal evidence the central interface of video self-evolution: the Questioner discovers evidence spans, generates evidence-grounded questions, and the Solver learns from pseudo supervision aligned with these spans.

Reinforcement learning for video reasoning. Video MLLMs have advanced general video understanding through improved spatio-temporal modeling, instruction following, and long-context video representation [45, 46]. Recent video reasoning and post-training methods further show that reinforcement learning can improve video MLLMs on video QA, temporal perception, and long-video reasoning [9–11, 20, 21, 47]. A common direction is to encourage models to reason over relevant temporal moments using temporal rewards, evidence-aware reasoning traces, or temporally grounded supervision [9, 10, 20, 22, 48–50]. These methods demonstrate the importance of temporal objectives for video post-training. Most existing video RL pipelines rely on externally constructed supervision, such as curated video QA data, cold-start reasoning traces, task-specific reward labels, expert demonstrations, or temporal annotations [9, 11, 21, 22]. In contrast, Video-Zero starts from unlabeled videos and automatically constructs evidence-grounded supervision through a Questioner–Solver loop. Thus, reinforcement learning is used not only to improve the Solver, but also to optimize the process by which useful video supervision is generated.

Temporal grounding and evidence-centric video understanding. Temporal grounding, moment localization, and intention forecasting require models to identify or reason about the video segments that support a language query or evolving human interactions [12, 29, 51–54]. These tasks highlight a core challenge in video understanding: correct answers and predictions often depend on when events happen, how interactions unfold over time, and what temporal evidence supports them, rather than on global video context alone. Recent video reasoning work also suggests that models may exploit shortcuts such as static frames, dataset priors, or language bias when temporal evidence is not explicitly enforced [9, 10, 20, 22]. Prior work typically uses temporal evidence as an annotation, an auxiliary training signal, or an evaluation target for a given question. Instead of treating evidence as a downstream target, Video-Zero uses evidence as the starting point for annotation-free self-evolution. The Questioner first identifies informative temporal spans from an unlabeled video and then generates question-answer pairs grounded in those spans. The Solver is trained to answer the question while aligning its prediction with pseudo-labeled evidence. This design turns temporal grounding from an external supervision source into the organizing principle of video self-evolution.

C More Implementation Details

This section summarizes the training data, iterative self-evolution pipeline, optimization settings, and evaluation protocol used in our experiments.

C.1 Training Data and Iterative Pipeline

Video-Zero performs five self-evolution iterations in the main experiments. Each iteration follows the path $S_t \to Q_{t+1} \to \mathcal{D}_{t+1} \to S_{t+1}$: the current Solver S_t is used to optimize the next Questioner Q_{t+1}, the optimized Questioner generates evidence-grounded supervision from unlabeled videos, and the retained pseudo-labeled data are used to train the next Solver S_{t+1}. Training data are regenerated at each iteration rather than reusing a fixed static dataset.

We use videos sampled from Open-o3-Video [22] as the unlabeled video source. Only raw videos are used; any associated questions, answers, or annotations are discarded. The main experiments use only 600 randomly sampled videos, while the data-scaling study uses 1K and 2K videos under the same self-evolution pipeline. For each video, the Questioner generates evidence-grounded supervision units containing a temporal evidence span, a question, and an answer. Invalid samples with missing fields, unparsable timestamps, empty answers, or invalid temporal ordering are discarded. For the main experiments, videos are sampled at 2 FPS with a maximum frame budget of 32 frames; frame-budget scaling experiments increase the maximum budget to 64 or 128 frames.

C.2 Pseudo Supervision and Rewards

For Solver training data construction, the frozen Questioner first generates candidate questions from the unlabeled videos. The current Solver then produces $n_{\text{pl}} = 10$ with-video rollouts for each question. We obtain the pseudo answer label $\hat{y}$ by majority vote and compute its support ratio s:

$$\hat{y} = \arg\max_{a \in \mathcal{A}} \sum_{j=1}^{n_{\text{pl}}} \mathbf{1}[f_j(v,q) = a], \qquad s = \frac{1}{n_{\text{pl}}} \max_{a \in \mathcal{A}} \sum_{j=1}^{n_{\text{pl}}} \mathbf{1}[f_j(v,q) = a].$$

Samples are retained only when $0.3 \le s \le 0.8$, filtering out questions that are either too unstable for reliable pseudo-labeling or too easy for the current Solver. For temporal supervision, we use the Solver rollouts whose final answer agrees with $\hat{y}$ and construct the consensus temporal target by taking the median start and end times across valid predicted spans:

$$\hat{\tau} = \left[\text{median}\{t_s^{(j)} : f_j(v,q) = \hat{y}\}, \text{median}\{t_e^{(j)} : f_j(v,q) = \hat{y}\}\right].$$

If the consensus span is invalid, we fall back to the Questioner-provided evidence span when available.

The Questioner reward follows the formulation in the main paper and combines format validity, learnability, video dependency, and evidence quality. In implementation, we normalize the utility components to comparable scales before aggregation:

$$r_Q = r_{\text{format}} + \tilde{r}_{\text{learn}} + \tilde{r}_{\text{dep}} + \tilde{r}_{\text{evid}}.$$

We set the format bonus to 0.1. The learnability and video-dependency terms are scaled by 0.5 and 0.3, respectively, before normalization, and the evidence-quality reward is clipped to a small range so that it does not dominate learnability and video dependency. The learnability term is $r_{\text{learn}} = \min(c, 1-c)$, where c is the majority-answer consistency under with-video Solver rollouts. The video-dependency term compares Solver agreement with the Questioner-provided answer under with-video and without-video conditions:

$$\Delta_{\text{video}} = \frac{1}{m}\sum_{j=1}^{m} \mathbf{1}[a_j^{\text{with}} = a^Q] - \frac{1}{m}\sum_{j=1}^{m} \mathbf{1}[a_j^{\text{without}} = a^Q], \qquad r_{\text{dep}} = \max(\Delta_{\text{video}}, 0).$$

Evidence-quality implementation. We compute three bounded scores from the Questioner's structured output. For nonempty evidence, the span score is $s_{\text{span}} = 0.10\,\mathbb{I}[N_s \ge 1] + 0.05\,\mathbb{I}[N_s \ge 2]$, where N_s is the number of parsed temporal spans and $\mathbb{I}[\cdot]$ denotes the indicator function. The event score is $s_{\text{event}} = 0.075\ \min(N_e, 2)$, where N_e counts text segments obtained by splitting the evidence description on periods and semicolons. The temporal-cue score is $s_{\text{temp}} = 0.10\ \min(20\rho_t, 1)$, with $\rho_t = N_t / \max(N_w, 1)$. Here, N_t counts case-insensitive matches from a fixed temporal-expression lexicon, such as *before*, *after*, *then*, and *during*, and N_w is the word count. The final reward is $r_{\text{evid}} = \text{clip}_{[-\beta,\beta]}(s_{\text{span}} + s_{\text{event}} + s_{\text{temp}} - \beta)$, with $\beta = 0.2$.

For Solver optimization, we use the evidence-aligned reward:

$$r_S = \mathbf{1}[\hat{a}^p = \hat{y}]\,(1 + \alpha r_{\text{align}}), \qquad r_{\text{align}} = \text{IoU}(\tau^p, \hat{\tau}) \prod_{b \in \{s,e\}} \left[1 - \frac{|t_b^p - \hat{t}_b|}{T}\right]_+.$$

Table 7: Key optimization settings used in Video-Zero.

Setting	Questioner	Solver
Optimization algorithm	GRPO	GRPO
Learning rate	1×10^{-6}	1×10^{-6}
Epochs per iteration	1	1
KL coefficient	0.01	0.01
Rollouts per prompt for GRPO	4	5
Rollouts for pseudo-labeling	–	10
Max response length	4096	2048
Training precision	bfloat16	bfloat16
Rollout engine	vLLM	vLLM

We set $\alpha = 0.5$. Invalid or unparsable Solver outputs receive zero answer correctness and therefore no temporal alignment reward.

C.3 Optimization Settings

Both the Questioner and Solver are optimized with GRPO. For both roles, we use a learning rate of 1×10^{-6}, one training epoch per iteration, and a KL coefficient of 0.01. The Questioner uses four sampled rollouts per prompt during optimization. The Solver uses five sampled rollouts per prompt during GRPO training and $n_{\text{pl}} = 10$ rollouts per generated question for pseudo-label construction. All training is conducted with bfloat16 precision and vLLM-based rollout generation on 8 NVIDIA H20 GPUs. Each self-evolution iteration takes approximately 7 hours. For fair comparison, Video-Zero and competing self-evolution baselines use the same hardware configuration, rollout budget, number of iterations, and optimization schedule unless otherwise specified.

C.4 Evaluation Protocol

We evaluate the 13 benchmarks using three types of evaluation protocols. For ActivityNet [28], ANet-RTL [27], Charades-STA [29], LongVideoReason [36], MMVU [37], VideoMathQA [38], and VideoMMMU [39], we follow the evaluation setting of OneThinker [11]. For LongVideoBench [31], MLVU [32], and Video-MME-Long [33], we use VLMEvalKit [55]. For NExT-GQA [30], LSDBench [34], and VideoNIAH [35], we follow the official evaluation protocols of the original benchmarks. All checkpoints and baselines use the same prompt template, frame budget, decoding setting, scoring function, and output parser within each benchmark.

For temporal grounding benchmarks, the model predicts a temporal span $p = [p_s, p_e]$ for each query. We compute temporal Intersection-over-Union with the ground-truth span $g = [g_s, g_e]$:

$$\text{tIoU}(p, g) = \frac{\max(0, \min(p_e, g_e) - \max(p_s, g_s))}{\max(p_e, g_e) - \min(p_s, g_s)}.$$

We report mean temporal IoU and recall at tIoU thresholds 0.3, 0.5, and 0.7. For NExT-GQA, a prediction is counted as correct only when both the answer is correct and the predicted temporal span reaches the specified tIoU threshold. Unparsable answers, missing spans, or invalid temporal spans are scored as incorrect.

We use deterministic decoding for the OneThinker-style and official benchmark evaluations, and follow the default VLMEvalKit configuration for LongVideoBench, MLVU, and Video-MME-Long. All decoding settings are fixed per benchmark and shared by all compared methods. Table 8 summarizes the split, number of evaluated samples, metric, and evaluation protocol for each benchmark. For VLMEvalKit benchmarks, LongVideoBench contains 1,337 validation questions, MLVU contains 2,592 questions, and Video-MME-Long is obtained by evaluating the full Video-MME benchmark and reporting the 900-question long-video subset.

Table 8: Evaluation protocol for the 13 benchmarks reported in the main paper.

Benchmark	Split / Subset	#Samples	Metric
Temporal Grounding			
ActivityNet	full	17,031	mIoU, R@0.3/0.5/0.7
ANet-RTL	full	228	mIoU, R@0.3/0.5/0.7
Charades-STA	full	3,720	mIoU, R@0.3/0.5/0.7
NExT-GQA	test	5,554	GQA@0.3, GQA@0.5
Long-video Understanding			
LongVideoBench	val	1,337	Accuracy
MLVU	test	2,592	Accuracy
Video-MME-Long	long subset	900	Accuracy
LSDBench	test	1,304	Accuracy
VideoNIAH	main-4try	1,350×4	4-try strict acc.
Video Reasoning			
LongVideoReason	full	1,000	Accuracy
MMVU	full	625	Accuracy
VideoMathQA	full	222	Accuracy
VideoMMMU	full	900	Accuracy

D More Experimental Results

This section provides additional experimental results that complement the main paper. We first report full iterative QA and reasoning results on Qwen3-VL-4B-Instruct and Qwen3-VL-8B-Instruct [24], including all intermediate checkpoints for Video-Zero and the competing self-evolution baselines VisPlay [13] and V-Zero [7]. We then present cross-architecture results on MiMo-SFT-7B [25] and InternVL-4B [26] to examine whether the evidence-centered self-evolution procedure transfers beyond the primary Qwen3-VL backbones. Finally, we provide full ablation results and qualitative examples to further analyze the contribution of each component and the evolution of Questioner-generated supervision.

D.1 Full Iterative Temporal Grounding Results

Table 9 reports the full temporal grounding trajectories on Qwen3-VL-4B and Qwen3-VL-8B. While Table 1 in the main paper provides a compact comparison using the Iter-4 checkpoints of VisPlay and V-Zero, this appendix table includes all intermediate checkpoints for Video-Zero and the self-evolution baselines across iterations.

Across both model scales, Video-Zero consistently improves temporal grounding over the base model. The gains are especially pronounced on the 4B backbone, where performance increases steadily across most metrics from early to later iterations. For the stronger 8B backbone, improvements are smaller but remain consistent, with the best checkpoints typically appearing in later iterations. Compared with competing self-evolution baselines, Video-Zero shows a clearer upward trajectory, supporting the effectiveness of evidence-centered supervision generation and temporal alignment for localization-oriented video understanding.

D.2 Full Iterative Video QA and Reasoning Results

Table 10 reports the full iterative results on long-video understanding and video reasoning benchmarks for Qwen3-VL-4B and Qwen3-VL-8B [24]. While Table 2 in the main paper provides a compact summary using the Iter-4 checkpoints of VisPlay, V-Zero, and Video-Zero, together with the peak Video-Zero checkpoint across iterations, this appendix table includes all intermediate checkpoints for

Table 9: Full iterative temporal grounding results on Qwen3-VL-4B and Qwen3-VL-8B. We report all intermediate checkpoints for Video-Zero, VisPlay, and V-Zero. The best and second-best results within each model block are highlighted in dark and light blue, respectively.

Methods	ANet-RTL				ActivityNet				Charades				NExT-GQA	
	mIoU	R@0.3	R@0.5	R@0.7	mIoU	R@0.3	R@0.5	R@0.7	mIoU	R@0.3	R@0.5	R@0.7	GQA@0.3	GQA@0.5
Qwen3-VL-4B-Instruct														
Base Model	16.86	22.70	15.72	8.73	22.86	30.94	19.04	11.29	37.52	58.90	36.12	16.23	23.09	15.40
VisPlay (Iter 1)	16.29	24.02	14.85	7.42	22.93	31.19	19.08	11.17	37.46	58.98	36.18	16.51	23.30	15.49
VisPlay (Iter 2)	16.44	22.27	14.85	8.30	22.76	30.84	18.82	11.12	37.49	58.66	36.18	16.37	23.36	15.72
VisPlay (Iter 3)	15.34	21.83	13.54	7.42	22.83	31.05	19.14	11.19	37.20	58.04	35.91	16.56	23.34	15.58
VisPlay (Iter 4)	16.58	23.14	13.97	8.73	22.89	30.98	19.12	11.31	37.59	58.82	36.59	16.42	23.14	15.33
VisPlay (Iter 5)	17.23	23.58	14.85	8.30	22.81	30.99	18.97	11.20	37.35	58.82	36.13	16.26	23.25	15.56
V-Zero (Iter 1)	16.43	23.14	14.41	8.73	23.07	31.24	19.17	11.40	37.46	58.66	36.13	16.69	23.07	15.52
V-Zero (Iter 2)	16.19	22.27	13.97	8.73	22.90	30.88	19.05	11.45	37.53	59.22	36.26	15.94	23.41	15.61
V-Zero (Iter 3)	15.88	22.27	13.54	7.86	22.63	30.55	18.85	11.28	37.12	57.98	35.51	16.53	23.50	15.61
V-Zero (Iter 4)	15.85	20.96	13.97	8.30	22.81	30.91	18.92	11.31	37.10	58.33	35.24	15.62	22.82	15.22
V-Zero (Iter 5)	16.65	22.71	15.72	8.73	23.12	31.50	19.22	11.41	37.55	58.84	36.59	16.29	22.64	15.05
Video-Zero (Iter 1)	20.86	28.82	20.09	8.73	24.23	32.97	20.22	12.04	38.75	60.59	36.91	17.45	31.66	20.80
Video-Zero (Iter 2)	23.81	33.19	22.27	11.35	27.30	37.48	23.16	13.66	40.85	64.19	39.81	18.60	36.59	22.80
Video-Zero (Iter 3)	27.75	39.30	28.82	13.97	30.29	41.82	26.15	15.50	41.66	64.54	41.40	19.60	38.14	23.25
Video-Zero (Iter 4)	29.75	44.98	28.82	13.54	32.12	44.76	27.67	16.24	41.85	65.11	40.81	20.03	38.29	23.00
Video-Zero (Iter 5)	29.34	42.79	27.95	13.54	32.59	45.35	28.01	16.38	41.48	64.62	39.92	19.95	37.76	22.82
Qwen3-VL-8B-Instruct														
Base Model	21.95	30.57	21.83	10.04	26.56	36.59	22.63	12.85	36.70	59.11	33.87	13.49	35.93	21.83
VisPlay (Iter 1)	22.17	31.00	21.40	11.35	26.66	36.68	23.13	12.88	36.66	58.92	33.71	13.41	36.00	22.04
VisPlay (Iter 2)	22.35	31.88	20.09	10.92	26.85	37.28	23.11	12.94	36.67	58.60	34.35	13.44	35.87	22.31
VisPlay (Iter 3)	21.21	31.00	18.34	9.17	26.57	36.85	22.99	12.73	36.58	58.76	33.44	13.52	36.16	22.22
VisPlay (Iter 4)	21.83	32.31	20.09	10.04	26.74	36.94	22.99	12.93	36.65	59.03	33.92	13.47	35.87	22.19
VisPlay (Iter 5)	21.95	32.31	19.21	9.61	26.75	36.94	23.18	13.03	36.70	59.11	34.01	13.33	36.03	22.20
V-Zero (Iter 1)	23.36	34.93	21.83	10.04	26.89	37.16	23.29	13.11	36.43	58.44	33.33	13.52	35.78	22.19
V-Zero (Iter 2)	23.05	31.88	19.21	10.04	26.76	36.88	23.01	13.09	36.66	59.11	33.52	13.31	35.51	22.19
V-Zero (Iter 3)	24.70	34.50	22.71	12.23	26.86	37.21	23.23	12.95	36.66	59.19	33.76	13.49	35.82	22.01
V-Zero (Iter 4)	22.77	32.31	20.52	9.17	26.69	36.79	23.05	13.00	36.43	58.63	33.66	13.41	35.73	22.11
V-Zero (Iter 5)	22.87	32.75	20.09	10.48	26.83	36.83	23.25	13.12	36.42	58.23	33.41	13.66	35.62	21.95
Video-Zero (Iter 1)	24.78	34.50	24.45	12.23	26.75	36.87	23.08	12.99	36.32	58.58	32.93	12.82	37.98	23.55
Video-Zero (Iter 2)	23.90	35.37	23.14	12.23	27.79	38.49	23.94	13.42	36.37	58.84	32.77	12.72	39.04	24.31
Video-Zero (Iter 3)	24.48	37.99	21.83	11.79	28.99	40.17	25.20	14.15	36.84	59.22	33.60	13.12	39.29	24.24
Video-Zero (Iter 4)	26.19	39.30	24.45	12.23	29.87	41.77	26.23	14.66	37.34	60.16	35.54	13.58	39.60	24.26
Video-Zero (Iter 5)	23.74	34.50	21.40	9.61	29.42	41.00	25.83	14.39	37.07	60.03	34.41	12.93	39.96	25.10

all self-evolution methods. These results complement the temporal grounding analysis by evaluating whether evidence-centered self-evolution transfers to broader video understanding and reasoning tasks.

Video-Zero generally improves over the base model across both model scales, with gains on benchmarks requiring long-context understanding, temporal evidence retrieval, or multi-step reasoning. Different benchmarks peak at different iterations, suggesting that the co-evolution process provides complementary benefits across task types rather than a single uniformly optimal checkpoint.

D.3 Cross-Architecture Results

Table 11 reports the full cross-architecture results on MiMo-SFT-7B [25] and InternVL3.5-4B [26]. While Table 3 in the main paper summarizes the compact cross-architecture comparison using the best checkpoint for each benchmark, this appendix table reports all intermediate checkpoints across the five self-evolution iterations. Across the two architectures, Video-Zero improves most video QA and reasoning benchmarks, with especially clear gains on reasoning-oriented tasks such as LongVideoReason, VideoMathQA, and VideoMMMU. The improvements are particularly strong for InternVL3.5-4B on reasoning benchmarks, while long-video understanding shows more backbone- and benchmark-dependent trends.

We additionally evaluate Video-Zero on ANet-RTL, ActivityNet, Charades, and NExT-GQA using MiMo-RL-7B and InternVL3.5-4B. For each backbone, the base model and Video-Zero use the same initialization.

Table 10: Full iterative video QA and reasoning results on Qwen3-VL-4B and Qwen3-VL-8B. We report all intermediate checkpoints for Video-Zero, VisPlay, and V-Zero. For clarity, dark and light blue highlight the best and second-best Video-Zero checkpoints within each model block, respectively.

Methods	Long-video Understanding					Video Reasoning			
	LongVideoBench	MLVU	Video-MME-L	LSDBench	VideoNIAH	LongVideoReason	MMVU	VideoMathQA	VideoMMMU
Qwen3-VL-4B-Instruct									
Base Model	57.70	62.60	51.60	54.68	45.78	69.90	64.32	19.05	57.56
VisPlay (Iter 1)	58.90	62.70	51.80	54.75	45.78	69.50	64.32	18.81	57.89
VisPlay (Iter 2)	59.60	63.50	51.10	54.91	45.33	70.30	63.52	20.71	56.78
VisPlay (Iter 3)	58.00	62.70	51.80	54.83	45.56	69.60	63.36	20.71	57.22
VisPlay (Iter 4)	59.20	63.50	53.20	54.83	45.26	69.60	63.36	18.33	57.44
VisPlay (Iter 5)	59.30	62.50	50.40	54.52	45.33	69.30	64.80	19.52	57.11
V-Zero (Iter 1)	58.60	62.80	51.90	54.60	45.63	69.90	64.16	20.48	57.67
V-Zero (Iter 2)	59.20	63.10	52.10	54.75	45.63	70.00	64.80	20.48	56.33
V-Zero (Iter 3)	59.60	63.40	52.80	54.60	45.63	68.40	64.32	19.52	57.67
V-Zero (Iter 4)	59.50	63.50	51.90	54.68	45.56	68.20	64.48	20.95	58.33
V-Zero (Iter 5)	59.20	62.80	52.00	54.83	45.48	69.60	64.16	20.48	57.89
Video-Zero (Iter 1)	58.20	62.90	51.80	54.98	45.70	68.00	64.32	19.76	57.11
Video-Zero (Iter 2)	59.40	63.00	51.30	55.21	45.70	69.90	66.56	19.05	58.89
Video-Zero (Iter 3)	59.10	62.80	52.80	55.29	45.78	69.90	63.20	21.19	59.00
Video-Zero (Iter 4)	59.10	63.40	53.60	56.52	46.37	70.30	64.96	21.67	61.56
Video-Zero (Iter 5)	59.60	62.90	52.20	56.67	46.07	70.50	65.60	22.62	60.44
Qwen3-VL-8B-Instruct									
Base Model	59.80	64.40	55.70	57.59	45.70	68.60	68.32	23.81	63.78
VisPlay (Iter 1)	59.80	64.10	55.90	57.36	45.85	68.50	68.16	24.76	64.11
VisPlay (Iter 2)	60.70	64.40	55.40	57.44	45.63	69.40	67.68	21.67	62.22
VisPlay (Iter 3)	60.00	64.00	56.60	57.21	45.70	68.40	67.68	23.33	63.67
VisPlay (Iter 4)	60.50	64.90	54.90	57.44	45.70	68.10	67.04	21.43	62.67
VisPlay (Iter 5)	60.40	64.40	56.40	57.59	45.85	68.20	67.36	22.38	64.22
V-Zero (Iter 1)	60.40	64.60	54.30	57.06	45.63	68.40	67.52	24.05	64.33
V-Zero (Iter 2)	59.80	63.90	55.00	57.29	45.70	69.30	68.48	22.86	64.22
V-Zero (Iter 3)	60.60	64.10	55.90	57.21	45.78	68.10	67.36	22.38	63.67
V-Zero (Iter 4)	60.10	64.20	56.00	56.90	46.07	69.20	68.80	22.14	63.22
V-Zero (Iter 5)	60.30	64.60	56.00	56.90	45.70	69.90	68.96	24.52	62.22
Video-Zero (Iter 1)	60.60	64.40	55.90	58.13	45.85	69.70	67.20	26.19	63.78
Video-Zero (Iter 2)	60.20	64.70	57.00	57.29	46.07	68.40	68.32	23.33	63.11
Video-Zero (Iter 3)	60.40	63.20	57.70	57.29	45.93	69.20	67.36	25.24	63.56
Video-Zero (Iter 4)	60.40	64.60	56.60	56.75	46.37	70.70	68.00	25.48	64.22
Video-Zero (Iter 5)	60.20	63.80	56.70	56.06	46.15	71.00	68.16	25.95	65.33

Table 12: **Cross-architecture temporal grounding results.** The first three benchmarks report mIoU; NExT-GQA reports GQA at IoU thresholds of 0.3 and 0.5.

Backbone	Method	ANet-RTL mIoU	ActivityNet mIoU	Charades mIoU	NExT-GQA @0.3	NExT-GQA @0.5
MiMo-RL-7B	Base	4.82	9.01	18.22	13.43	6.99
	Video-Zero	**5.07**	**10.32**	**20.04**	**14.03**	**7.40**
InternVL3.5-4B	Base	1.86	14.26	14.93	5.80	2.52
	Video-Zero	**7.05**	**14.56**	**16.73**	**6.50**	**2.56**

D.4 Full Ablation Results

The main paper reports ablation results at Iter 4, which serves as a unified comparison point for all ablated variants. Table 4 summarizes the category-level averages, where the grounding average is computed over all temporal grounding metrics, and the video QA and reasoning averages are computed over the corresponding benchmark groups used in the main experiments. The overall average is then computed from these category-level averages. Tables 13 and 14 provide the detailed Iter 4 results: Table 13 reports full temporal grounding metrics, while Table 14 reports representative long-video understanding and video reasoning benchmarks.

Table 11: Cross-architecture results on long-video understanding and video reasoning benchmarks. Panel (a) reports results on MiMo-SFT-7B, and panel (b) reports results on InternVL3.5-4B. Columns v1–v5 denote Video-Zero checkpoints across self-evolution iterations. Bold numbers indicate the best checkpoint for each benchmark. Δ denotes the maximum improvement over the base model.

(a) MiMo-SFT-7B

Benchmark	Base	v1	v2	v3	v4	v5	Δ
Long-video Understanding							
LongVideoBench	47.60	50.00	**53.00**	49.10	50.20	51.20	+5.40
MLVU	40.40	41.90	43.00	43.60	43.30	**43.70**	+3.30
Video-MME-Long	33.10	34.10	**35.90**	34.40	34.70	34.00	+2.80
LSDBench	51.92	54.45	54.29	**55.83**	55.67	55.14	+3.91
VideoNIAH	46.67	**46.81**	46.74	45.26	45.41	44.96	+0.14
Video Reasoning							
LongVideoReason	69.10	72.60	**73.10**	72.50	71.80	68.50	+4.00
MMVU	69.60	**70.08**	**70.08**	68.16	68.00	68.16	+0.48
VideoMathQA	29.76	30.48	**32.62**	29.52	**32.62**	27.86	+2.86
VideoMMMU	67.33	67.56	67.44	**68.33**	67.78	67.56	+1.00

(b) InternVL3.5-4B

Benchmark	Base	v1	v2	v3	v4	v5	Δ
Long-video Understanding							
LongVideoBench	60.30	61.00	60.40	**61.20**	60.70	60.90	+0.90
MLVU	**62.40**	62.00	61.80	61.80	61.90	61.50	-0.40
Video-MME-Long	52.40	53.30	53.30	53.30	53.20	**53.70**	+1.30
LSDBench	**60.35**	57.52	57.21	57.06	56.98	57.36	-2.83
VideoNIAH	**38.81**	36.89	36.89	36.81	36.67	36.44	-1.92
Video Reasoning							
LongVideoReason	68.00	**72.00**	**72.00**	71.40	71.60	71.70	+4.00
MMVU	54.08	57.60	58.40	**58.88**	58.72	58.40	+4.80
VideoMathQA	12.62	17.86	21.90	23.10	23.33	**23.81**	+11.19
VideoMMMU	44.22	**53.11**	51.78	52.56	52.00	52.56	+8.89

Table 13: Full temporal grounding ablation results at Iter 4 on Qwen3-VL-4B. All variants are evaluated at the same iteration for a unified ablation protocol.

Method	ANet-RTL				ActivityNet				Charades				NExT-GQA	
	mIoU	R@0.3	R@0.5	R@0.7	mIoU	R@0.3	R@0.5	R@0.7	mIoU	R@0.3	R@0.5	R@0.7	GQA@0.3	GQA@0.5
Video-Zero	29.75	44.98	28.82	13.54	32.12	44.76	27.67	16.24	41.85	65.11	40.81	20.03	38.29	23.00
Description-only Questioner	17.70	24.45	16.59	9.61	23.59	32.18	19.78	11.72	37.85	59.57	35.97	16.91	20.94	13.56
w/o Video Dependency	23.19	34.06	21.40	10.92	30.32	41.87	26.24	15.28	40.00	61.80	38.68	17.88	38.23	23.84
w/o Evidence Quality	28.24	41.48	27.07	14.41	31.28	43.66	27.22	15.75	38.38	59.41	36.21	17.10	38.59	23.61
w/o Learnability	22.80	31.88	21.40	9.17	29.64	41.25	24.90	14.97	39.33	60.48	37.18	18.39	36.50	22.13
w/o Temporal Reward	14.72	19.65	13.10	7.42	20.91	28.53	17.17	10.18	36.88	58.74	35.30	15.70	26.44	17.32
Frozen Questioner	26.95	39.30	25.33	11.79	30.28	42.12	26.05	15.40	40.24	61.94	38.25	19.87	37.42	23.18

Table 14: Representative long-video understanding and video reasoning ablation results at Iter 4 on Qwen3-VL-4B. All variants are evaluated at the same iteration; category-level averages in the main paper are computed over the full benchmark groups.

Method	Long-video Understanding		Video Reasoning			
	Video-MME-L	LSDBench	LongVideoReason	MMVU	VideoMathQA	VideoMMMU
Video-Zero	53.60	56.52	70.30	64.96	21.67	61.56
Description-only Questioner	52.70	55.29	69.40	65.60	22.14	57.89
w/o Video Dependency	51.90	55.21	71.20	66.08	20.71	58.78
w/o Evidence Quality	52.00	56.98	70.90	64.96	20.24	58.78
w/o Learnability	52.90	54.52	70.40	64.96	21.43	60.11
w/o Temporal Reward	52.30	55.52	68.90	62.56	18.10	57.00
Frozen Questioner	51.40	56.21	69.00	63.68	21.90	57.67

Source of Solver supervision. The supervision-source ablations in Table 15 compare three answer–evidence target combinations. The full method uses the Solver-consensus answer $\hat{y}$ and temporal target $\hat{\tau}$ defined in Section 3.3. Q-Ans uses the Questioner-generated answer a^Q in place of $\hat{y}$ while retaining the Solver-consensus temporal target $\hat{\tau}$. Q-Ans & Q-Span replaces both targets with the Questioner's answer and proposed evidence, denoted by (a^Q, e^Q).

Table 15: **Source of Solver supervision.** Q-Ans replaces the consensus answer with the Questioner's answer; Q-Ans & Q-Span also replaces the temporal target with Questioner evidence. Higher scores are better.

Supervision	Grounding Avg.	QA & Reasoning Avg.	Overall Avg.
Solver consensus $(\hat{y}, \hat{\tau})$	**33.02**	**55.28**	**48.43**
Q-Ans $(a^Q, \hat{\tau})$	23.29	54.14	44.65
Q-Ans & Q-Span (a^Q, e^Q)	24.05	53.91	44.72

D.5 Evidence Analysis Protocol

We provide additional details for the evidence analysis in Figure 2(a)(b). The goal is to examine whether questions generated by Video-Zero depend more strongly on temporally localized video evidence than those generated by V-Zero [7]. To avoid circular evaluation, we use Qwen3-VL-235B-A22B-Thinking-FP8 [24] as an external Oracle model only for annotation, and use a fixed base Solver for all evaluations.

Video dependency. Figure 2(a) evaluates whether generated questions require the video input rather than being answerable from language priors alone. For each generated question, we evaluate the fixed base Solver under two conditions: with the video and without the video. We report the with-video accuracy and the video-dependency gap, defined as the difference between with-video and without-video accuracy. A larger gap indicates that the generated question depends more strongly on visual evidence in the video. From Iteration 1 to Iteration 4, Video-Zero increases both the with-video accuracy and the dependency gap, while V-Zero shows weaker or decreasing trends. This suggests that evidence-centered question generation produces supervision that becomes more video-grounded over iterations.

Temporal evidence sensitivity. Figure 2(b) further evaluates whether the generated questions depend on a specific temporal span. For each generated question, the Oracle first produces a unified reference answer and then annotates the most critical continuous temporal span for answering the question. All methods and iterations are annotated by the same Oracle. The Oracle-annotated span is treated as the key temporal evidence for the question. Given this key span, we construct four controlled frame conditions for Solver evaluation: *Full*, which uses the full video; *Only-key*, which uses only frames inside the key span; *Mask-key*, which removes frames inside the key span and keeps the remaining frames; and *Random*, which uses an equal-length random span that does not overlap with the key span when possible. The fixed base Solver is evaluated under all four conditions, and its predictions are compared against the Oracle reference answer.

We focus on two diagnostic metrics that quantify dependency on the Oracle-annotated key span:

$$\text{Key Necessity} = \text{Acc}_{\text{full}} - \text{Acc}_{\text{mask}}, \qquad \text{Key Specificity} = \text{Acc}_{\text{key}} - \text{Acc}_{\text{rand}}. \tag{7}$$

Key Necessity measures whether the key span is indispensable by comparing full-video accuracy with accuracy after masking the key span. Key Specificity measures whether the key span is more informative than an arbitrary span by comparing the key span with an equal-length random span. Higher values indicate that the generated question depends more strongly on localized temporal evidence rather than generic visual context.

Table 16 reports the Iteration-4 comparison. Video-Zero achieves a Key Necessity of 27.96% and a Key Specificity of 25.81%, compared with 15.05% and 11.83% for V-Zero. This indicates that Video-Zero generates questions whose answers depend more strongly on specific temporal evidence.

Table 16: **Iteration-4 evidence analysis.** We evaluate a fixed base Solver under controlled frame conditions derived from Oracle-annotated key temporal spans. Video-Zero shows higher Key Necessity and Key Specificity than V-Zero.

Method	N	$\mathbf{Acc}_{\mathrm{full}}$	$\mathbf{Acc}_{\mathrm{key}}$	$\mathbf{Acc}_{\mathrm{mask}}$	$\mathbf{Acc}_{\mathrm{rand}}$	Key Nec. / Key Spec.
Video-Zero	93	64.52	54.84	36.56	29.03	27.96 / 25.81
V-Zero	93	55.91	53.76	40.86	41.94	15.05 / 11.83

Table 17: Frame-budget robustness on temporal grounding benchmarks. We report results under 64-frame and 128-frame max-frame settings on Qwen3-VL-4B, with all settings using 2 FPS sampling. Iter 4 follows the unified comparison point used in the main paper, and Δ_{peak} denotes the maximum improvement over the corresponding base model across iterations.

Frames	Checkpoint	ANet-RTL mIoU	ActivityNet mIoU	Charades mIoU	NExT-GQA@0.3	NExT-GQA@0.5
64f	Base	17.89	24.41	38.77	23.36	15.52
64f	Iter 4	31.39	32.36	40.81	37.82	23.18
64f	Δ_{peak}	+13.50	+9.94	+2.04	+15.12	+8.12
128f	Base	17.24	24.58	38.70	23.28	15.31
128f	Iter 4	20.86	28.22	28.73	35.22	21.95
128f	Δ_{peak}	+6.37	+5.43	+0.00	+12.18	+6.93

D.6 Robustness, Scaling, and Qualitative Analysis

We provide additional analyses to examine the robustness and behavior of Video-Zero beyond the main experimental setting. Specifically, we evaluate Video-Zero across different frame budgets, training data scales, and random seeds, and further include qualitative examples that illustrate how Questioner-generated supervision evolves over self-evolution iterations.

D.6.1 Frame-Budget Robustness

We provide detailed results for the frame-budget robustness analysis in Figure 5. All settings use the same 2 FPS sampling rate and differ only in the maximum number of sampled frames. In addition to the main 32-frame setting, we evaluate Qwen3-VL-4B under 64-frame and 128-frame budgets on temporal grounding benchmarks. Table 17 reports results for these larger frame budgets. For compactness, we report the corresponding base model, the Iter 4 checkpoint used for the unified comparison in the main paper, and the peak improvement over the corresponding base model across iterations.

Overall, Video-Zero remains effective under larger frame budgets. The improvements are especially clear on ANet-RTL, ActivityNet, and NExT-GQA, showing that evidence-centered supervision remains beneficial when more visual context is provided. Gains vary by benchmark, with mixed results on Charades.

D.6.2 Effect of Training Video Duration

Training-video composition and frame budget. The original training pool contains 600 unlabeled videos shorter than three minutes, with a median duration of 36.6 seconds. LongMix replaces 200 of these short videos with 200 unlabeled videos lasting 5–20 minutes, retaining a total of 600 training videos. Both training variants and all evaluations use a maximum of 64 frames. The comparison keeps the backbone, reward design, training procedure, and checkpoint selection fixed between the two training variants.

LongMix improves three of four long-video benchmarks and all four reasoning benchmarks over short-only training. Grounding changes are mixed: Charades remains nearly unchanged, NExT-GQA@0.3 decreases, and NExT-GQA@0.5 improves. All three grounding metrics remain above the base model.

Table 18: **Full LongMix results under a fixed 64-frame budget.** Δ is LongMix minus short-only Video-Zero. Bold numbers indicate the best score among the three settings for each benchmark.

Benchmark	Base	Video-Zero 600 short	Video-Zero LongMix	Δ
Temporal grounding				
Charades mIoU	38.77	**40.81**	40.79	-0.02
NExT-GQA@0.3	23.36	**37.82**	36.63	-1.19
NExT-GQA@0.5	15.52	23.18	**23.54**	$+0.36$
Long-video understanding				
MLVU	66.40	66.80	**67.20**	$+0.40$
Video-MME-L	52.20	53.11	**54.33**	$+1.22$
LSDBench	55.06	**55.90**	55.83	-0.07
VideoNIAH	57.25	57.55	**57.62**	$+0.07$
Video reasoning				
LongVideoReason	70.30	71.30	**72.00**	$+0.70$
MMVU	62.72	65.44	**67.20**	$+1.76$
VideoMathQA	18.57	20.48	**21.90**	$+1.42$
VideoMMMU	58.56	58.78	**60.22**	$+1.44$

Table 19: Training data scaling on temporal grounding benchmarks. We report peak results across self-evolution iterations for each unlabeled video scale on Qwen3-VL-4B. Avg. is computed over the five reported metrics.

Data Scale	ANet-RTL mIoU	ActivityNet mIoU	Charades mIoU	NExT-GQA@0.3	NExT-GQA@0.5	Avg.
600	29.75	32.59	41.85	38.29	23.25	33.15
1K	31.42	33.76	41.92	39.15	24.47	34.14
2K	**32.58**	**38.21**	**46.54**	**40.12**	**24.76**	**36.44**

D.6.3 Training Data Scaling

We study how Video-Zero scales with the amount of unlabeled video data used for self-evolution. In addition to the 600-video setting used in the main experiments, we evaluate larger unlabeled video pools with 1K and 2K sampled videos while keeping the same Questioner–Solver pipeline, reward design, and optimization procedure. Tables 19 and 20 summarize the peak results across self-evolution iterations for each data scale.

Overall, increasing the unlabeled video pool improves temporal grounding, with the average grounding score increasing from 33.15 at 600 videos to 36.44 at 2K videos. The gains are especially clear on ANet-RTL, ActivityNet, and Charades, suggesting that a larger raw video pool helps the Questioner discover more diverse temporal evidence. For long-video understanding and video reasoning, scaling brings additional gains at 2K, although the trend is less monotonic than temporal grounding. This indicates that additional unlabeled videos mainly strengthen temporal evidence learning, while downstream QA and reasoning improvements also depend on the match between generated supervision and benchmark requirements.

D.6.4 Random Seed Robustness

We evaluate the robustness of Video-Zero under different random seeds. In addition to the main run, we train two additional runs with different random seeds while keeping the same 600 unlabeled videos, training pipeline, reward design, and evaluation protocol. Table 21 summarizes the peak iteration-level averages across the three runs.

For temporal grounding, we first average all reported metrics from ANet-RTL, ActivityNet, Charades, and NExT-GQA within each iteration, and then report the best iteration-level average for each seed. For QA, we average the five long-video understanding benchmarks and four video reasoning benchmarks within each iteration, and then report the best iteration-level QA average for each seed. Across three seeds, Video-Zero achieves peak iteration-level averages of 32.76 ± 0.69 on temporal

Table 20: Training data scaling on long-video understanding and video reasoning. We report peak category averages across self-evolution iterations for each unlabeled video scale on Qwen3-VL-4B.

Data Scale	Long-video Avg.	Reasoning Avg.
600	55.88	55.31
1K	55.62	54.59
2K	**56.87**	**55.63**

Table 21: Random seed robustness on the Qwen3-VL-4B backbone. We combine the main run with two additional seeds and report peak iteration-level category averages across iterations.

Seed	Temporal Grounding Avg.	QA Avg.
Main	33.36	55.28
123	32.92	54.25
456	32.00	53.90
Mean $\pm$ Std.	32.76 ± 0.69	54.48 ± 0.72

grounding and 54.48 ± 0.72 on QA. These results show that Video-Zero remains stable across random seeds for both temporal grounding and broader video QA.

D.6.5 Qualitative Analysis of Questioner-generated Supervision

We provide qualitative case studies from two complementary perspectives. Figures 6 and 7 compare questions generated by different self-evolution methods on the same videos. V-Zero tends to produce coarse event-level or commonsense questions, while VisPlay often focuses on surface-level attributes or fails to produce a valid formatted question in some cases. By contrast, Video-Zero generates questions that are more tightly coupled with temporal evidence, visual transitions, and fine-grained scene changes.

Figures 8 and 9 further illustrate how the Video-Zero Questioner evolves across self-evolution iterations on the same video. Early iterations tend to produce coarse scene-level questions or short-horizon temporal queries, whereas later iterations generate more fine-grained, interaction-aware, and temporally structured supervision. These examples illustrate the video-dependence and temporal specificity of the generated supervision.

D.7 Limitations, Broader Impacts, and Future Work

Video-Zero demonstrates that video VLMs can self-evolve from unlabeled videos by organizing generated supervision around temporally localized evidence. Our results across multiple model backbones show that Video-Zero consistently improves temporal grounding, long-video understanding, and video reasoning performance. A positive impact of this work is that it reduces the need for manually annotated video QA and temporal grounding data, potentially making video model development more scalable and accessible.

However, the primary limitation of this work is that we were unable to validate this scaling trend on larger video VLMs, such as 32B, 72B, or 100B+ models, due to the prohibitive computational cost. This leaves the full scaling behavior of evidence-centered self-evolving training as an important direction for future work. Another limitation is that Video-Zero relies on model-generated questions, answers, and temporal evidence, which may inherit biases or errors from the underlying video VLMs. For sensitive video domains, additional caution and domain-specific validation would be important. Future work will explore larger backbones, stronger filtering and verification, and responsible release with documentation, intended-use guidance, and misuse restrictions.

Qualitative Comparison of Evidence-Centered Question Generation



Caption: A Christmas-themed lounge scene shifts from people filming and observing to a small live holiday performance.

V-Zero: Event Recognition ☹

Question: Which of the following best describes the activity taking place in the video?
Answer : A holiday musical performance.
Type: Event Recognition

VisPlay: Coarse Attribute Prediction ☹

Question: How many people are visible in the scene at the peak of the performance?
Answer : 6.
Type: Coarse Attribute Prediction

Video-Zero: Temporal Focus Shift ☺

Iter 5 : Temporal Focus Shift
Question: Which sequence correctly describes the transition observed between the initial moments and the main performance scene?
Answer : An individual filming moves away, revealing a group of musicians about to perform.

Figure 6: **Qualitative comparison on a holiday performance video.** Video-Zero generates a temporally grounded question about the transition into the main performance, while V-Zero and VisPlay focus on coarser event or attribute cues.

Qualitative Comparison of Evidence-Centered Question Generation



Caption: A glacier ice cave scene shifts from a blue-lit interior to the rocky exterior and back to detailed ice formations.

V-Zero: Generic Cave Knowledge ☹

Question: If one were to explore an ice cave near a glacier, what would most likely be observed?
Answer : Intricate formations of ice and rock illuminated by natural light.
Type: Generic Cave Knowledge

Video-Zero: Spatial Perspective Reasoning ☺

Question: What does the camera's shift between the ice cave interior and the outside glacier landscape reveal?
Answer : It reveals a transition from sunlit ice formations to rocky glacier terrain with a person nearby, before returning to detailed ice textures.

Figure 7: **Qualitative comparison on a glacier ice-cave video.** Video-Zero captures the camera's spatial perspective shift across cave interior, glacier exterior, and ice formations, while V-Zero asks a generic commonsense question.

Evolution Across Iterations on the Same Video: Family Meal Interaction (video length: 150s)



Caption: A family dining scene gradually shifts from coarse scene understanding to fine-grained reasoning about mother–child interaction and feeding behavior.

V-Zero

Question: What can be inferred about the environment?
Answer : A home environment with informal, relaxed atmosphere.
Type: Static scene / background cue ☹

Video-Zero

Iter 1 : Scene Focus 😐
Question: How does the video focus change over time?
Answer : It shifts from a group dining scene to a close-up of the woman eating.
Solver score: 0.9

Iter 3 : Action Interaction 😐
Question: How does the adult remain engaged with the meal while responding to the child's presence?
Answer : The adult continues eating while intermittently responding to the child, who watches and attempts to imitate the action.
Solver score: 0.8

Iter 5 : State Transition ☺
Question: At what point does the interaction shift from parallel eating to direct feeding?
Answer : After beginning to eat herself, the woman starts offering food to the child with a spoon.
Solver score: 0.6

Figure 8: **Evolution of Video-Zero Questioner on a family meal video.** Across iterations, the questions evolve from coarse scene focus to fine-grained reasoning about mother–child interaction and direct feeding.

Evolution Across Iterations on the Same Video: Bullfighting (video length: 79s)



Caption: A rodeo scene where a bull is released, throws off the rider, and then charges around the arena while cowboy tries to evade and control it.

V-Zero

Question: What is most likely happening during this event?
Answer : A bull riding competition.
Type: Scene recognition / event category ☹

Video-Zero

Iter 1 : Coarse temporal localization 😐
Question: When does the bull first engage with the man after being released?
Answer : Between 0.0s and 2.0s
Solver score: 0.9

Iter 3 : Local consequence reasoning 😐
Question: What happens after the rider's initial mounting attempt?
Answer : The rider is thrown off shortly after mounting and then tries to get back up.
Solver score: 0.7

Iter 5 : Multi-agent Tracking and Cross-segment Comparison ☺
Question: How does the man's behavior relative to the bull differ between the early charge and a later confrontation?
Answer : He runs ahead while keeping distance during the first charge, followed by close maneuvers leading to a chase threat.
Solver score: 0.6

Figure 9: **Evolution of Video-Zero Questioner on a bullfighting video.** Across iterations, the questions progress from coarse temporal localization to consequence reasoning and multi-agent tracking.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes].

   Justification: The claims accurately describe Video-Zero as an annotation-free, evidence-centered Questioner–Solver co-evolution framework and are supported by experiments across temporal grounding, long-video understanding, and video reasoning benchmarks.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes].

   Justification: The paper includes a dedicated limitations discussion.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A].

Justification: The paper does not present theorem-style theoretical results; the equations define the proposed training objectives, rewards, and pseudo-supervision construction.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes].

   Justification: The paper and appendix provide the necessary implementation, training, and evaluation details to reproduce the main experimental results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes].

Justification: Raw videos used in this study are publicly available and can be easily accessed online. The code for reproducing our results will be released upon acceptance of the paper by the conference.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes].

Justification: The paper and appendix specify the training settings, evaluation splits, metrics, optimizer, and key hyperparameters.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes].

Justification: The appendix reports multi-seed results and robustness analyses to assess the stability of the observed gains.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes].

   Justification: The appendix reports the GPU setup, training precision, and approximate cost for the main experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes].

   Justification: The research follows the NeurIPS Code of Ethics and does not involve human-subject experiments or sensitive personal data collection.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes].

    Justification: The paper discusses both the potential benefits and risks of applying self-evolution to video model training.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A].

    Justification: Our framework do not have such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes].

    Justification: The paper credits the datasets, benchmarks, model backbones, and software tools used in the experiments through citations and follows their stated terms of use where applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes].

    Justification: The released code and checkpoints are documented with intended use, reproduction instructions, limitations, and licensing information.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A].

    Justification: The paper does not involve crowdsourcing or human-subject experiments.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A].

    Justification: The paper does not involve human-subject experiments, so IRB approval is not applicable.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A].
    item[] Justification: The core method development in this research does not involve LLMs as an important,original components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.