Structure Over Scale: Learning Visual Reasoning from Pedagogical Video

Bishoy Galoaa*
Northeastern University
galoaa.b@northeastern.edu

Xiangyu Bai*
Northeastern University
bai.xiang@northeastern.edu

Sarah Ostadabbas
Northeastern University
s.ostadabbas@northeastern.edu

Abstract

State-of-the-art vision-language models (VLMs) score impressively on video benchmarks yet stumble on basic visual reasoning tasks involving spatial relations, navigation, and object selection that a preschooler solves easily. We hypothesize that the explicit pedagogical structure, specifically the *context-question-pause-answer* cycles embedded in children's educational video, provides naturally co-aligned reasoning traces: temporally synchronized visual cues, questions, and answers that emerge only from deliberate pedagogical authoring and cannot be practically reconstructed through manual annotation at scale. To test this, we introduce **SoSVQA** (Structure over Scale Visual Question Answering), a unified benchmark of 10K question-answer pairs automatically extracted from *Dora the Explorer* (DoraVQA) and *Mickey Mouse Clubhouse* (ClubHVQA) with precise timestamp alignment, and fine-tune VLMs using Group Relative Policy Optimization (GRPO) to leverage the clear correctness signals and structured reasoning traces inherent in educational content. Holding each backbone and its pretraining fixed, adding 10K pedagogical QA pairs from 78 hours of video yields gains of up to 19.7 points on NExT-QA and 10.6 on Video-MME. A matched-size generic QA control does not reproduce these gains. Additional experiments on InternVL3-8B and Qwen3.5-4B support transfer beyond the original backbones. Performance is competitive on these aligned benchmarks, while gains on fine-grained motion and temporal benchmarks are small or absent. Code is available at ostadabbas/SoS.

1 Introduction

Despite strong video-benchmark performance [34, 33], vision-language models still struggle with spatial relations, navigation, and object selection [8, 27]. These failures motivate examining how training data connects language to visual evidence. We study pedagogical video, where questions, relevant frames, and explicit answers recur in a consistent temporal structure. Our hypothesis is that this organization supplies useful supervision beyond simply adding more generic video-text pairs; it does not imply that data scale is unimportant or that small fine-tuned models uniformly outperform larger systems.

Children's educational television offers this uniquely structured alternative. Programs such as *Dora the Explorer* [28] and *Mickey Mouse Clubhouse* [10] are explicitly designed around a consistent *context-question-pause-answer* pedagogical loop: each episode establishes visual context, poses direct spatial questions to the viewer, and then pauses (visually "hinting" at the answer by overlaying

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

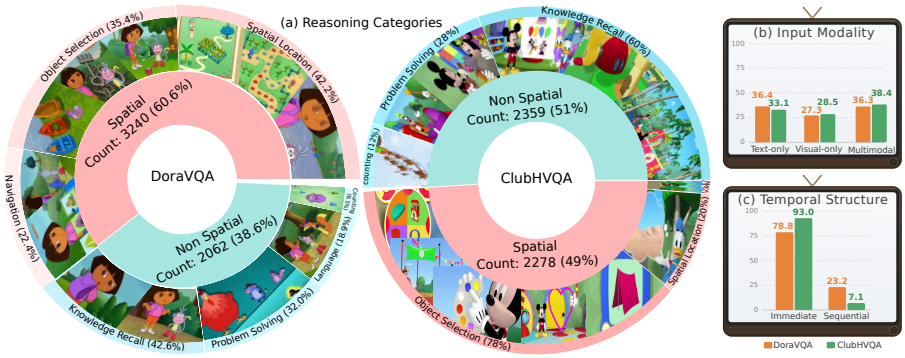


Figure 1: **SoSVQA** (Structure over Scale Visual Question Answering) dataset composition across two subsets: DoraVQA and ClubHVQA. **(a) Reasoning categories:** DoraVQA is dominated by spatial-reasoning tasks, whereas ClubHVQA emphasizes object-selection and knowledge-recall tasks. **(b) Input modality:** Both datasets exhibit similar input modality distributions. **(c) Temporal structure:** ClubHVQA predominantly requires immediate reasoning, whereas DoraVQA includes a substantially larger portion of sequential-reasoning questions.

directional arrows, zooms, and character gestures that isolate the exact evidence needed to answer) before revealing an unambiguous answer with verbal explanation. This pause segment functions as an implicit reasoning trace, encoding two distinct learning signals that we later isolate experimentally (Section B): a spatial attention signal (where to look, recoverable through the visual hints) and an entity grounding signal (what it is called). The structure recurs systematically across every entry, generating 15–20 explicit question-answer pairs with clear correctness signals per episode.

This pedagogical design produces measurable learning gains in children [17, 2, 3], with controlled studies showing that preschoolers who watch *Dora the Explorer* score significantly higher on direction/position concepts and spatial ability tasks than children who do not [3], confirming that the pause structure carries genuine supervisory signal. We hypothesize that the same structure that supports human learning can serve as an efficient and underexplored training signal for vision language models. This observation points to a broader scalable data strategy: rather than blindly rely on increasing data scale; we need data that knows how to *teach*. Fortunately, the world contains a large supply of pedagogically structured content that has never been used this way: children's television, Khan Academy lectures, PhET simulations, and similar resources all embed explicit teaching signals in their design. To test this hypothesis, we introduce **SoSVQA** (Structure over Scale Visual Question Answering), a unified benchmark of 10K question-answer pairs, containing two sub-datasets extracted from children's educational television: **DoraVQA**, comprising 5,348 question-answer pairs from 8 seasons (96 episodes) of *Dora the Explorer*, and **ClubHVQA** (ClubhouseVQA), comprising 4,652 question-answer pairs from 4 seasons (132 episodes) of *Mickey Mouse Clubhouse*. Each example preserves the show's pedagogical structure (context, question, pause, and answer) with corresponding video timestamps and pause-based visual context extracted via an LLM agent.

To test whether structured supervision transfers beyond its source content, we fine-tune Qwen2-VL [29] and Qwen3-VL [4] using Group Relative Policy Optimization (GRPO) [22], which naturally aligns with educational content: answers are objectively correct or incorrect, and the show's explanations provide implicit reasoning cues without manual annotation. Rather than adapting models to specific content, our goal is to improve spatial reasoning abilities using structured learning format already embedded in the data. We evaluate along three axes: (i) **in-domain** performance on DoraVQA and ClubHVQA, (ii) **cross-show** transfer between the two subsets to test whether learning generalizes across pedagogical content rather than memorizing show-specific patterns, and (iii) **cross-benchmark** transfer to Video-MME [7], NExT-QA [33], MVBench [15], MotionBench [12], and TVBench [6]. We train on open-ended answer generation but evaluate on multiple-choice benchmarks throughout, a deliberate format mismatch that distinguishes transferable reasoning from answer-format memorization. Our central comparison holds the backbone and training recipe fixed: the matched-size generic-data control in Table 5(b), together with within-SoSVQA structure ablations, tests whether extra QA supervision alone explains the gains. We additionally probe whether the show's *pause hints* are recoverable by prompting alone or require structured supervision (Section B), separating spatial attention from entity grounding as distinct learning signals.

Our contribution can be listed as follows:

- We show that the *context-question-pause-answer* structure of children's educational television serves as implicit reasoning supervision for VLMs, yielding cross-benchmark spatial reasoning improvements without large-scale data.
- We present an automated pipeline that extracts self-supervised training data from pedagogically structured video, preserving the correctness signals embedded in its reasoning structure without manual annotation. We validate the pipeline across two shows with distinct reasoning profiles, and additionally release manually annotated bounding-box hints for the pause-segment visual cues to support future spatial-grounding work.
- We introduce **SoSVQA**, a benchmark of 10K question-answer pairs automatically extracted from *Dora the Explorer* and *Mickey Mouse Clubhouse* with precise timestamp alignment. SoSVQA is the first video question-answering datasets extracted from children's educational television, with temporal alignment between visual frames, questions, context, and answers.

2 Related Work

VLMs and Video Understanding. Recent work has scaled VLMs to capture temporal dynamics beyond images. Video-LLaVA [16] demonstrated that joint training on videos and images improves both modalities through unified visual representations; InternVideo2 [31] expanded video encoders using progressive training and masked video modeling; LLaVA-Video [37] introduced dense frame sampling to capture fine-grained temporal dynamics. However, these approaches train predominantly on web videos [18, 1, 13], which contain diverse content but lack explicit pedagogical structure. While these videos display spatial relationships and temporal sequences, they rarely verbalize spatial concepts explicitly or provide structured question-answer pairs that ground visual patterns in linguistic descriptions. Our work differs by leveraging educational content where spatial reasoning is explicitly taught through scaffolded question-answer sequences with clear correctness signals.

Spatial Reasoning in VLMs. Despite achieving strong performance on existing video benchmarks [33, 34], VLMs exhibit systematic failures on spatial reasoning tasks. Fu et al. [8] reveals that VLMs achieve only half of human accuracy with noticeable failures on spatial reasoning tasks. Tong et al. [27] demonstrated that current VLMs are barely above random chance on tasks requiring precise spatial understanding. These failures extend to fundamental spatial primitives: models struggle with counting objects beyond small numbers [19], fail at compositional spatial relationships [26], and fail to reason about object positions across video frames [21]. Importantly, these deficiencies persist even in simple visual scenes with minimal occlusion and clear object boundaries, indicating that the limitation is not perceptual but conceptual: models lack the structured spatial reasoning that children acquire through explicit instruction.

Educational and Instructional Video as Training Data. Instructional videos have been explored as training data for procedure understanding and action recognition. Miech et al. [18] demonstrated that instructional videos provide natural language supervision for video-text embeddings. However, the narrations describe actions chain rather than teaching concepts; The CrossTask [38] and COIN [25] datasets extract procedural steps from instructional videos but focus on temporal ordering rather than spatial reasoning. Although these datasets compose of structured contents, they are not explicit question-answer structure that "teaches" spatial reasoning. Closest to our work, Anderson and Burns [2] analyzed the pedagogical structure of children's educational television, documenting how the question-pause-answer format creates structured learning opportunities. We build on this insight by extracting the explicit question-pause-answer pairs from educational content as direct supervision signals, leveraging the pedagogical design for model training rather than treating video as passive observation. Adjacent work has trained VLMs on infant egocentric video [30], but no prior work has used the pedagogical structure of children's television as a direct training signal for VLMs.

Reinforcement Learning for VLM Fine-tuning. Group Relative Policy Optimization (GRPO) [22] has emerged as an efficient alternative to Proximal policy optimization (PPO) [20] for reasoning tasks: DeepSeek-R1 [11] demonstrated that GRPO with rule-based rewards achieves reasoning capabilities competitive with proprietary models. Recent work has successfully applied GRPO to vision-language tasks: VLM-R1 [23] showed that GRPO fine-tuning achieves superior out-of-distribution generalization compared to SFT on visual reasoning tasks, while CrowdVLM-R1 [32]

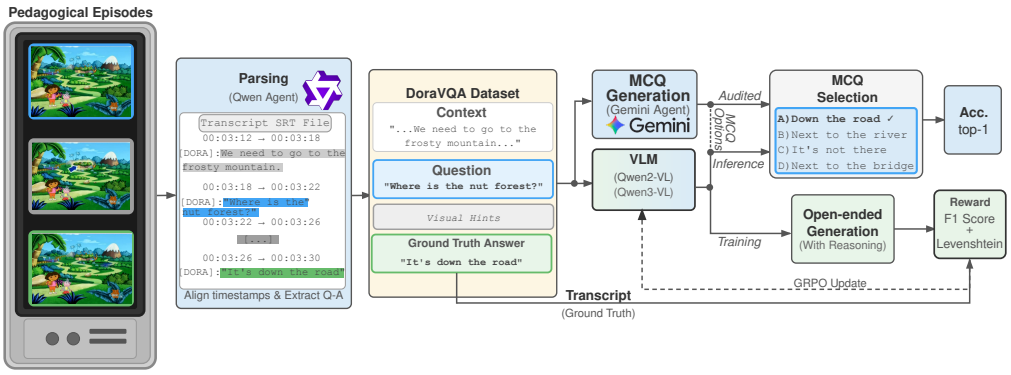


Figure 2: **SoSVQA pipeline overview.** We extract question-answer pairs from pedagogical episodes by parsing SubRip Text (SRT) transcript files with a Qwen agent, aligning timestamps to identify the show's pedagogical *context-question-pause-answer* structure. Each question is paired with its surrounding context window, the ground truth answer, and multiple-choice distractors (MCQs) that follows. During training, we fine-tune Qwen models using GRPO on open-ended generation; during inference, the MCQ options are provided to our fine-tuned models, creating a deliberate train-test format mismatch (open-ended $\rightarrow$ MCQ) that evaluates transferable reasoning.

developed specialized reward functions for counting tasks in crowded scenes. These works primarily design reward models for general-purpose instruction following. In contrast, we apply GRPO to pedagogically-structured video content where the reward signal is implicit in the educational design, where ground-truth answers are directly in its structure. This eliminates the need for external reward model while providing clear correctness signals, creating a self-supervised RL training environment where the pedagogical content naturally defines the learning objectives. Motion-o [9] complements this direction by making object trajectories explicit in video reasoning.

Our work is positioned at the intersection of these research directions: we address video-language model spatial reasoning failures by training on content that explicitly teaches spatial concepts through structured question-answer pairs. Unlike general instructional videos that show only procedures, educational television teaches reasoning through direct instruction, repetition, and scaffolded examples. We leverage this pedagogical structure with GRPO fine-tuning, using the natural correctness signals from educational content as supervision without requiring manual annotation or reward engineering.

3 Method

3.1 SoSVQA: Structure and Formatting

We operationalize the *context-question-pause-answer* structure of educational television as a multimodal training example (Figure 2). Each example in the SoSVQA dataset $\mathcal{D}$ is formatted as $x = \{I, T, Q, a^*\} \in \mathcal{D}$, where I represents the visual frames sampled before and at the question timestamp, T is the transcript providing narrative context within a fixed time window before the question, Q is the explicit question and a^* is ground truth answer. We treat I and T as separate input streams rather than collapsing them, allowing the model to draw on visual scene evidence and linguistic context independently before fusion. The ground-truth answer a^* is always either directly stated in the transcript or deducible in a single inferential step, providing an unambiguous reward target without manual annotation. The model receives x as a structured multimodal prompt (visual frames embedded alongside a text sequence concatenating the system instruction, transcript context, and question) and generates a distribution over answer tokens that is trained toward a^* via the reward signal described in Section 3.2.

3.2 Group Relative Policy Optimization (GRPO)

We fine-tune using GRPO [22], which optimizes a policy π_θ without a separate value network by computing advantages relative to a group of sampled outputs. GRPO is well-suited here because educational content provides objectively correct answers, eliminating the need for complex reward modeling. For each input $x \in \mathcal{D}$, the policy samples K candidate answers $\{a_1, \ldots, a_K\}$. Each

Algorithm 1 GRPO Training with Context-Question-Pause-Answer Structure

Require: Dataset $\mathcal{D}$ of x=$\{I, T, Q, a^*\}$; policy π_θ; reference π_{ref}; group size K; weights α, β, λ; learning rate η

for each mini-batch $\mathcal{B} \sim \mathcal{D}$ **do**
  for all $x \in \mathcal{B}$ **do**
    Sample $\{a_i\}_{i=1}^{K} \sim \pi_\theta(\cdot \mid I, T, Q)$
    $r_i \leftarrow \alpha\,\text{F1}(a_i, a^*) + \beta\Big(1 - \frac{\text{lev}(a_i, a^*)}{\max(|a_i|, |a^*|)}\Big)$
    $A_i \leftarrow r_i - \frac{1}{K}\sum_j r_j$
  end for
  $\theta \leftarrow \theta - \eta\,\nabla_\theta \sum_{x,i}[-A_i \log \pi_\theta(a_i \mid x) + \lambda\, D_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}})]$
end for

candidate is scored against the ground-truth a^* via:

$$r(a_i, a^*) = \alpha\,\text{F1}(a_i, a^*) + \beta\left(1 - \frac{\text{lev}(a_i, a^*)}{\max(|a_i|, |a^*|)}\right), \tag{1}$$

where $\text{F1}(a_i, a^*)$ measures token-level semantic overlap, $\text{lev}(a_i, a^*)$ is the Levenshtein edit distance normalized by the length of the longer string $\max(|a_i|, |a^*|)$, and $\alpha = 0.3$, $\beta = 0.7$ are fixed weights selected by empirical validation (see Appendix A), with $\alpha + \beta = 1$. The two terms are complementary: F1 tolerates minor surface variation while normalized Levenshtein penalizes character-level errors and near-miss re-orderings. This yields a stable continuous signal; exact matches score 1.0, semantically correct answers with minor variation score 0.65–0.95, and incorrect answers score near zero.

For a group of K sampled candidates with rewards $\{r(a_1, a^*), \ldots, r(a_K, a^*)\}$, the advantage for candidate a_i is computed relative to the group mean:

$$A(a_i, a^*) = r(a_i, a^*) - \frac{1}{K}\sum_{j=1}^{K} r(a_j, a^*), \tag{2}$$

so that the policy is updated toward outputs that outperform the group average rather than toward an absolute value target. The training objective is:

$$\mathcal{L}(\theta) = -\mathbb{E}_{x\sim\mathcal{D},\, a_i\sim\pi_\theta(\cdot|x)}\left[A(a_i, a^*) \log \pi_\theta(a_i|x)\right]. \tag{3}$$

Critically, the reward in Eq. 1 is computed against free-form a^* during training, while evaluation uses multiple-choice prompting. This deliberate format mismatch tests whether gains reflect transferable reasoning rather than answer-format memorization.

3.3 GRPO Training

We use standard GRPO to study the supervision supplied by the data, rather than to claim an algorithmic contribution. The GRPO training procedure for the *context-question-pause-answer* structure is presented in Algorithm 1 and illustrated in Figure 2. For each multimodal input $x \in \mathcal{D}$, the policy π_θ generates K candidate answers $\{a_1, \ldots, a_K\}$ via temperature sampling (temperature $= 1.0$, top-$p = 0.9$). Each candidate a_i is evaluated against the ground-truth answer a^* using the reward function $r(a_i, a^*)$ defined in Section 3.2, and group-relative advantages $A(a_i, a^*)$ are computed over the K generations. The policy parameters θ are updated by minimizing $\mathcal{L}(\theta)$ to maximize the expected group-relative advantage. To stabilize training, a KL divergence penalty is added between the updated policy π_θ and a frozen reference policy π_{ref}:

$$\mathcal{L}_{\text{KL}}(\theta) = \mathcal{L}(\theta) + \lambda\,\mathbb{E}_{x\sim\mathcal{D}}\left[D_{\text{KL}}(\pi_\theta(\cdot|x) \,\|\, \pi_{\text{ref}}(\cdot|x))\right], \tag{4}$$

where $\lambda = 0.01$ is the KL coefficient. Unless otherwise noted, models are trained jointly on both SoSVQA splits; single-show training is used only in the cross-show ablation (Table 5). Training hyperparameters (learning rate $\eta = 1 \times 10^{-4}$, reward scaling factor $s = 2.0$, and group size $K = 8$) are selected via sequential sweeps on held-out validation data (Appendix A).

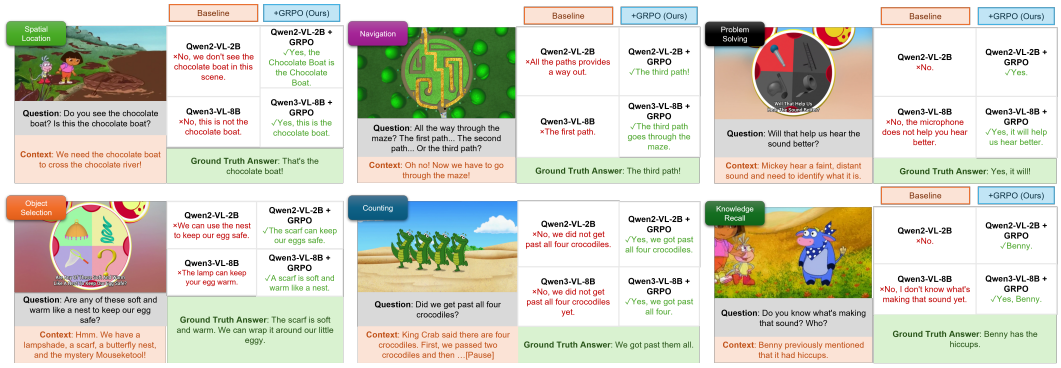


Figure 3: **Qualitative comparison on some samples of challenging spatial reasoning tasks.** Our GRPO-finetuned models (blue) outperform all baselines (orange) across every reasoning category. While baseline models frequently fail or hallucinate, our models consistently produce accurate and concise answers. See Appendix C and supplementary material for more examples.

4 Experiments and Results

We evaluate our approach along three axes designed to test the structure-over-scale hypothesis. First, *in-domain* performance on both SoSVQA splits (DoraVQA and ClubHVQA) tests whether the *context-question-pause-answer* structure was learned at all, and contamination analysis verifies that improvements come from visual grounding rather than pre-training memorization. Second, *cross-show* transfer between DoraVQA and ClubHVQA tests whether learning generalizes across pedagogical content rather than memorizing show-specific patterns. Third, *cross-benchmark* transfer to five external video-understanding benchmarks tests whether the learning generalizes beyond children's educational content, and ablations isolate which components drive the gains.

4.1 Datasets

We evaluate on in-domain SoSVQA (DoraVQA and ClubHVQA) and a broad set of out-of-domain external benchmarks selected to test the transferability of spatial reasoning improvements. Transfer varies by benchmark; we report both improvements and regressions.

SoSVQA. The two SoSVQA subsets contain a combined total of 10,000 QA pairs and follow a consistent *context–question–pause–answer* pedagogical structure with clear ground-truth supervision, but exhibit complementary reasoning profiles (Figure 1): DoraVQA (5,348 pairs) is spatial-heavy (60.6%) with strong navigation demands, while ClubHVQA (4,652 pairs) is balanced between spatial and non-spatial tasks (49%/51%) and dominated by object selection (78%). We do not redistribute video frames or audio; we release only episode identifiers, timestamps, QA annotations, MCQ choices, transcript span indices, the annotation/alignment pipeline, and manually annotated bounding boxes. Users must obtain audiovisual content independently from the rights holders under the hosting platform's terms (Appendix E). Multiple-choice options are generated from the free-form gold answers via Gemini 2.5 Flash and manually audited to prevent surface-form shortcuts (Appendix F).

Video-MME [7] evaluates multimodal LLMs across short, medium, and long videos using visual, subtitle, and audio cues; we report the short partition. **NExT-QA** [33] tests causal and temporal reasoning through "why" and "how" questions. **MVBench** [15] covers 20 temporal video-understanding tasks designed to require multi-frame reasoning. **MotionBench** [12] evaluates fine-grained motion comprehension through six motion-oriented question categories. **TVBench** [6] isolates temporal reasoning by removing the spatial, textual, and world-knowledge shortcuts present in earlier benchmarks.

4.2 In-Domain Performance on SoSVQA

Table 1 reports per-category performance on both DoraVQA and ClubHVQA, testing whether the *context-question-pause-answer* structure was learned. GRPO fine-tuning produces large, structurally-aligned gains on the two categories most directly targeted by the show's pedagogical pause: spatial reasoning improves by +11 to +13 points across all DoraVQA rows, and navigation gains as much as +23.25 (Qwen2-VL-2B) and +15.29 (Qwen3-VL-8B) on ClubHVQA, reflecting

Table 1: In-domain top-1 accuracy (%) on DoraVQA and ClubHVQA. **+ GRPO** uses both training splits. Gray rows are proprietary models; underlined scores are surpassed by Qwen3-VL-8B + GRPO. ‡Separate frontier subset: seed 0, 534 questions per show, four frames (one Gemini DoraVQA parse failure); category definitions are in Appendix L. Additional-backbone transfer results are in Appendix J.

	Overall		**Spatial**		**Obj./Count.**		**Navigation**		**Knowledge**	
Model										
Gemini-3.0-Flash	76.10	69.16	67.83	62.18	87.10	73.10	65.38	65.91	88.52	68.57
Gemini-3.0-Flash (low)	75.84	63.96	66.81	54.98	88.17	68.52	69.23	61.36	86.89	60.83
GPT-4V	67.79	68.20	60.34	64.10	75.27	69.90	69.23	65.91	68.85	66.15
Gemini-2.5-Pro	64.41	62.69	62.62	52.92	72.83	70.53	60.78	65.91	71.93	60.00
Gemini-3.6-Flash‡	79.92	77.72	71.98	75.76	83.87	83.08	67.31	71.43	95.00	86.55
GPT-5-mini‡	66.29	56.55	56.90	62.12	80.65	63.85	55.77	57.14	73.77	56.30
LLaVA-Video-7B	55.41	46.46	54.02	45.10	62.22	54.52	66.00	47.62	67.31	48.66
Video-LLaVA-7B	37.82	39.19	47.49	41.48	60.66	42.78	46.34	56.76	32.43	40.68
InternVideo2.5-8B	57.68	50.10	46.55	49.52	63.04	55.87	59.62	36.36	68.85	53.14
Qwen2-VL-2B	40.29	31.43	45.36	33.99	37.61	36.10	39.03	23.26	39.95	32.87
Qwen2-VL-2B + GRPO	**48.32**	**36.68**	**58.86**	**45.21**	**44.71**	**42.38**	**44.30**	**46.51**	**45.97**	**38.07**
Qwen2-VL-7B	52.21	45.09	54.14	44.55	51.45	49.51	41.97	40.91	48.77	48.05
Qwen2-VL-7B + GRPO	**57.68**	**50.10**	**67.25**	**50.97**	**64.37**	**58.98**	**47.80**	**50.00**	**67.25**	**54.49**
Qwen3-VL-8B	55.24	50.67	57.17	49.20	53.58	58.62	**48.96**	45.45	50.32	50.50
Qwen3-VL-8B + GRPO	**62.98**	**56.81**	**69.65**	**59.16**	**64.59**	**64.87**	48.33	**60.74**	**55.12**	**57.48**
InternVL3-8B	55.21	53.71	55.82	52.56	46.02	53.55	**54.78**	**68.18**	53.62	54.71
InternVL3-8B + GRPO	**62.41**	**57.61**	**66.36**	**55.77**	**60.27**	**59.84**	52.95	65.12	**55.92**	**55.67**
Qwen3.5-4B	46.17	39.06	46.10	39.09	39.55	39.94	**46.45**	29.27	42.70	37.91
Qwen3.5-4B + GRPO	**52.29**	**50.96**	**55.59**	**54.42**	**50.35**	**52.04**	39.31	**52.38**	**47.66**	**50.29**

the dataset's visually-driven structure. Object selection and counting improve consistently across subsets (+6 to +13 points), and knowledge recall transfers strongly, peaking at +18.48 for Qwen2-VL-7B on DoraVQA. These per-category gains aggregate into consistent overall improvements of +5.47 to +8.03 on DoraVQA and +5.01 to +6.14 on ClubHVQA, with larger relative gains on smaller models. The lone regression is Qwen3-VL-8B's navigation on DoraVQA (-0.63), whose baseline is the strongest among the three original fine-tuned backbones in this category. InternVL3-8B and Qwen3.5-4B also improve both subsets (Table 1), extending the evidence to five backbones; transfer results appear in Appendix J. The frontier rows in Table 1 use a separate subset, and static-puzzle results appear in Appendix L. Figure 3 shows representative qualitative improvements.

4.3 Cross-Benchmark Transfer

Table 2 places the controlled before/after comparisons in the context of model scale. Each fine-tuned model shares its baseline's pretraining and adds only 10K SoSVQA pairs. The original three backbones gain 12.46–19.70 points on NExT-QA and 1.75–10.64 on Video-MME. InternVL3-8B gains 1.66 and 6.78 points, respectively; Qwen3.5-4B gains 6.47 and 1.67. These results support transfer across model families and generations. Their partial evaluations are reported in Appendix J, Table 14; MVBench, MotionBench, and TVBench were not evaluated for these runs. Other open and proprietary models lead on several benchmarks. Open-ended training transfers to multiple-choice evaluation, reducing the likelihood of answer-format memorization.

The largest gains occur on NExT-QA and Video-MME, consistent with benefits from learning to connect a question to relevant visual and temporal evidence. On MVBench, MotionBench, and TVBench, gains are small or absent. This pattern bounds the claim: pedagogical supervision improves particular reasoning behaviors rather than video understanding uniformly. Across four Qwen3-VL-8B GRPO runs, Video-MME reaches 80.30 ± 0.56 and NExT-QA 81.40 ± 0.94 (mean

Table 2: Cross-benchmark transfer with model/data scale as context. Each GRPO row is compared with its own pretrained baseline; additional fine-tuning uses 10K SoSVQA pairs (78 hours). ▢ proprietary models; ▢ Qwen3-VL-8B fine-tuning; ▢ our minimal fine-tuning addition. Green indicates improvement over baseline. † indicates zero-shot evaluation; * indicates training set seen during pretraining. Our GRPO models share the same pre-training corpus as their baselines but additionally fine-tune jointly on the 10K pedagogically-structured QA pairs from SoSVQA (78 hours of educational video).

Model	Params	Scale	V-MME	NExT-QA	MVBench	MotionBench	TVBench
GPT-4V†	∼1.8T	∼10T tokens	59.9	68.2	43.50	58.82	34.00
Gemini-2.5-Pro†	–	–	85.2	74.6	45.50	66.30	46.50
Gemini-3.0-Flash†	–	–	**86.9**	80.4	65.50	70.40	38.25
InternVideo2.5-8B	8B	16M clips	65.4	71.5*	75.7	45.56	48.80
LLaVA-Video-7B	7B	1.3M videos	63.3	**83.2***	69.7	48.29	42.30
Video-LLaVA-7B	7B	760K videos	45.3	52.1*	40.95	35.72	36.30
Qwen2-VL-2B (baseline)	2B	1.2T tokens	50.10	52.60	58.99	47.0	39.85
Qwen2-VL-2B + GRPO†	2B	1.2T + **10K QA**	**60.74** (+10.64)	**72.11** (+19.51)	**59.94** (+0.95)	51.92 (+4.92)	40.05 (+0.20)
Qwen2-VL-7B (baseline)	7B	1.2T tokens	67.50	67.0	**65.49**	52.0	**40.75**
Qwen2-VL-7B + GRPO†	7B	1.2T + **10K QA**	**69.25** (+1.75)	**79.46** (+12.46)	64.33 (-1.16)	**54.01** (+2.01)	40.60 (-0.15)
Qwen3-VL-8B (baseline)	8B	1.0T tokens	71.40	62.1	**68.70**	**61.12**	46.53
Qwen3-VL-8B + GRPO†	8B	1.0T + **10K QA**	**80.00** (+8.60)	**81.80**† (+19.70)	67.59 (-1.11)	61.03 (-0.09)	47.48 (+0.95)

Table 3: Textual recall and visual-grounding probes. **(a)** LLM-only probes disable the vision encoder: MCQ tests episode-specific recall; F1 and containment score open-ended recall. Containment measures whether the gold answer occurs in the output. **(b)** Qwen3-VL-8B on the rebuttal evaluation subset, with identical questions and transcript context in the blank/frame conditions. Δ vision is the within-model frame-minus-blank accuracy in percentage points. The 25% MCQ chance level is not a baseline for comparing the different probe configurations.

(a) LLM backbone (text only)

	MCQ		F1		Containment	
Model						
Qwen2-VL-2B	34.60	19.20	5.23	4.32	22.01	19.14
Qwen2-VL-7B	36.37	21.00	5.46	9.49	23.74	23.65
Qwen3-VL-8B	39.23	27.38	12.07	11.57	24.64	16.22

(b) Qwen3-VL-8B (subset)

Model	Dataset	Blank	Frames	Δ vision
Base		44.54	58.08	+13.54
Base		48.87	50.45	+1.58
+ GRPO		45.23	67.98	+22.75
+ GRPO		46.43	56.81	+10.38

± sample standard deviation), versus baselines of 71.40 and 62.10; variability is small relative to the gains. Appendix K reports individual runs and the far-domain CLEVRER evaluation, where changes are small and mixed. On the synthetic-physics probe, Qwen3-VL-8B improves slightly on all three metrics, whereas the smaller backbones show both gains and declines. The static-image probes in Appendix L further test transfer across input format: all three backbones improve on VisualPuzzles, while SMART-101 gains concentrate in the 8B model.

4.4 Contamination Analysis

A natural concern is that gains reflect textual knowledge of the shows rather than visual grounding. Table 3 separates two configurations. The LLM-only MCQ and open-ended recall probes in panel (a) disable the vision encoder and test pretraining recall. The full-VLM blank-image probe in panel (b) retains transcript context, so its accuracy also reflects context exploitation. These probes are not directly comparable, and blank-image accuracy is above the 25% MCQ chance level, for Qwen3-VL-8B.

The relevant test compares blank images and episode frames within each model. For Qwen3-VL-8B, the complete 2×2 in Table 3(b) shows the vision gap increasing from 13.54 to 22.75 points on DoraVQA and from 1.58 to 10.38 on ClubHVQA after GRPO. Blank-image accuracy changes by only +0.69 and -2.44 points, respectively. Thus the small base-model ClubHVQA gap does not imply redundant frames: fine-tuning enables greater use of visual evidence. These within-model comparisons support improved grounding without claiming that contamination is absent.

A complementary counterfactual keeps the question, transcript, and choices fixed but swaps in frames from another episode on a seeded 534-question subset (Table 4). Jointly trained Qwen2-VL-

Table 4: Frame-swap counterfactual on a seeded 534-question subset. Qwen2-VL-2B + GRPO is evaluated with intact or different-episode frames, with question, transcript, and choices fixed. Deltas are percentage points; visual-required deltas use the flagged question subset. These are paired subset controls, not full-test scores.

Training	Evaluation	Intact	Swapped	Δ overall	Δ visual-required
Joint	DoraVQA	37.83	33.45	-4.38	-5.3
Joint	ClubHVQA	33.33	27.53	-5.81	-6.9
Dora-only	DoraVQA	48.69	45.51	-3.18	-1.8
Dora-only	ClubHVQA	30.34	26.40	-3.94	-5.2

Table 5: Training-data ablations. **(a)** Cross-show transfer with Qwen2-VL-2B: a model trained on one SoSVQA split is evaluated on both, plus three external benchmarks. **(b)** Generic (NExT-OE) training, 10K QA, across three backbones; the comparison with SoSVQA-joint training in Tables 1 and 2 tests whether matched-size generic QA exposure reproduces the gains. ▢ shaded cells denote in-domain evaluation. * indicates eval set seen during fine-tuning.

(a) Cross-show transfer (Qwen2-VL-2B)						**(b) Generic training (NExT-OE, 10K QA)**				
Train	**Method**	**In-Domain**		**Cross-Benchmark**		**Model**	**In-Domain**		**Cross-Benchmark**	
		🐢	🐠	V-MME	NExT		🐢	🐠	V-MME	NExT
–	Baseline	41.36	31.43	50.10	52.60	Qwen2-VL-2B	40.11	30.67	51.55	72.00*
🐢	SFT	54.30	30.52	53.33	66.63	Qwen2-VL-7B	53.55	44.30	65.33	82.33*
🐢	**GRPO**	**55.11**	**54.45**	**55.40**	**72.86**	Qwen3-VL-8B	**56.20**	**50.75**	**74.44**	**83.33***
🐠	**GRPO**	**67.04**	**59.92**	**77.53**	**80.25**					

2B loses 4.38 points on DoraVQA and 5.81 on ClubHVQA, including 6.9 points on visual-required ClubHVQA questions. These paired controls use the same run and evaluation harness; their subset accuracies should not be compared directly with full-test scores. Together, the blank-image and frame-swap tests support a contribution from correctly aligned visual evidence.

4.5 Ablation Studies

We conduct ablation studies on training strategies, input modalities, and pedagogical structure using a Qwen2-VL-2B model fine-tuned for 100 epochs on DoraVQA, evaluating performance on 10% randomly sampled test subsets of Video-MME and NExT-QA.

Training Method. Table 5(a) compares SFT and GRPO across shows. SFT matches GRPO in-domain on DoraVQA (54.30 vs. 55.11), but remains near the ClubHVQA baseline (30.52 vs. 31.43). GRPO transfers in both directions: DoraVQA→ClubHVQA gains 23.02 points, and ClubHVQA→DoraVQA gains 25.68. The ClubHVQA-trained model also outperforms the DoraVQA-trained model on DoraVQA itself (67.04 vs. 55.11). This baseline-free comparison supports learning shared structure rather than simply benefiting from familiarity with the target show's content.

Generic Training Control. To isolate whether GRPO itself drives the cross-benchmark gains, Table 5(b) trains the same backbones with GRPO on a generic 10K-pair open-ended VQA set (NExT-OE) of matched size. Despite NExT-OE training-set leakage (marked*), SoSVQA scores barely move from baseline (Qwen3-VL-8B: 56.20/50.75 vs. 55.24/50.67), and transfer to V-MME is also weaker than SoSVQA-joint training (74.44 vs. 80.00). Extra generic QA training therefore does not reproduce the SoSVQA gains; within-source structure ablations provide complementary evidence for the role of alignment.

Reward-Weight Sensitivity. Across five Qwen3-VL-8B F1/Levenshtein ratios (1:9–9:1), every setting improves both transfer benchmarks over baseline (Appendix A). The selected 3:7 ratio performs best (80.00/81.80); alternatives span 74.78–77.61 on Video-MME and 72.25–78.33 on NExT-QA. Thus gains persist across tested weights, although their magnitude varies. The results support the chosen edit-distance emphasis without establishing a universal optimum.

Context Modalities. Table 6(a) shows the model learns more effectively from transcripts than visual frames (50.95 vs. 43.42 on DoraVQA), consistent with the text-vision asymmetry reported in the Qwen2 technical report [29]. Combining modalities compounds the gain on both shows, with the

Table 6: Ablation study. **(a)** Input modality ablation over visual frames (V), transcript (T), and pedagogical context (C). **(b)** Pedagogical structure ablation with all modalities present. A vertical rule separates in-domain benchmarks (DoraVQA, ClubHVQA) from out-of-domain ones (V-MME and NExT-QA). Best results per section in **bold**. †Shared full-model row.

(a) Modality Ablation							(b) Structure Ablation				
V	T	C			V-MME	NExT	Structure			V-MME	NExT
✓	✗	✗	43.42	26.34	60.74	67.91	Random Clips	49.81	41.63	51.70	70.08
✗	✓	✗	50.95	31.88	60.00	70.48	No Pause	45.31	39.43	51.11	72.00
✓	✓	✗	54.72	39.77	61.11	70.83					
✓	✓	✓	**55.11**	**54.45**	**62.20**	**72.86**	Full†	**55.11**	**54.45**	**62.20**	**72.86**

Table 7: Effect of pause-based visual hints on the blue-arrow subset of DoraVQA (frames with arrows ≥ 1, Season 2 excluded). The four conditions vary the spatial signal available at inference: **No hint** (none), **Self-localized** (model's own bbox), **Prompted hint** (textual instruction only), **Oracle box** (ground-truth bbox). *Loc.* reports mean IoU and is only meaningful for *Self-localized*. **Bold**: best overall per column. Underline: best per model.

Model	Condition	Overall	Spatial	Nav.	Know.	Prob. solving	Loc. (mIoU)
Qwen3-VL-8B	No hint	52.24	58.17	35.38	56.10	66.67	–
	Self-localized	53.61	55.83	39.52	64.86	73.61	0.76
	Prompted hint	55.94	64.59	31.54	57.14	79.17	–
	Oracle box	55.25	61.24	45.60	46.15	67.61	–
Qwen3-VL-8B + GRPO	No hint	63.19	70.48	34.62	52.38	86.11	–
	Self-localized	66.98	**80.00**	36.92	**69.05**	83.33	**2.57**
	Prompted hint	**67.61**	77.14	**46.92**	52.38	84.72	–
	Oracle box	**67.61**	77.62	43.85	64.29	**87.50**	–

multimodal jump particularly pronounced on ClubHVQA (V+T: 39.77 vs. T-only: 31.88), reflecting its higher proportion of visually-grounded object-selection tasks.

Pedagogical Structure. Removing the pause segment drops DoraVQA accuracy by 9.80 points, and substituting random (non-pause) clips loses 5.30 points (Table 6(b)). The temporal alignment between question and pause-segment visual evidence is itself a learning signal: the same frames presented without pause alignment recover only part of the gain. To further isolate the two learning signals encoded in the pause—*spatial attention* (where to look) and *entity grounding* (what it is called)—we evaluate four conditions on a held-out blue-arrow subset of DoraVQA, varying how directly the arrow's location is made available to the model (Table 7). Prompting recovers most of the spatial-attention signal but cannot bridge from location to a named entity; only GRPO supervision closes that gap. The pattern holds across model scales up to 72B (Appendix B, Tables 11 and 12), where larger untuned models partially recover spatial attention through scale alone but never match the GRPO-finetuned 8B on entity grounding.

What the pause contributes. The controls suggest that pause alignment associates the question and correctness signal with relevant visual evidence. This may aid temporal credit assignment, but we do not measure reward gradients over visual tokens. Together with generic-data and cross-show controls, the results support aligned supervision without establishing claims about all pedagogical content.

5 Conclusion

Standard GRPO on 10K SoSVQA pairs improves five backbones on both source datasets and the primary transfer benchmarks. Generic-data, structure, and visual counterfactual controls support aligned pedagogical supervision. Mixed far-domain results limit the scope of transfer; broader pedagogical sources remain future work.

Funding and Competing Interests

This project received no funding. The authors declare no competing financial interests.

References

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. In *arXiv preprint arXiv:1609.08675*, 2016.

[2] Daniel R Anderson, Jennings Bryant, Alice Wilder, Angela Santomero, Marsha Williams, and Alisha M Crawley. Researching blue's clues: Viewing behavior and impact. *Media Psychology*, 2(2):179–194, 2000. doi: 10.1207/S1532785XMEP0202_4.

[3] Duriye Esra Angın. The effects of Dora the Explorer on preschool children's spatial concept acquisition and spatial ability. *European Scientific Journal*, 13(1):39–53, 2017. doi: 10.19044/esj.2017.v13n1p39.

[4] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[5] Anoop Cherian, Kuan-Chuan Peng, Suhas Lohit, Kevin Smith, and Joshua B. Tenenbaum. Are deep neural networks SMARTer than second graders? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. URL `https://arxiv.org/abs/2212.09993`.

[6] Daniel Cores, Michael Dorkenwald, Manuel Mucientes, Cees G. M. Snoek, and Yuki M. Asano. TVBench: Redesigning video-language evaluation. *arXiv preprint arXiv:2410.07752*, 2024.

[7] 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.

[8] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *ECCV*, 2024.

[9] Bishoy Galoaa, Shayda Moezzi, Xiangyu Bai, and Sarah Ostadabbas. Motion-o: Trajectory-grounded video reasoning. *arXiv preprint arXiv:2603.18856*, 2026. URL `https://arxiv.org/abs/2603.18856`.

[10] Bobs Gannaway. Mickey mouse clubhouse. Television series. Disney Channel (Playhouse Disney/Disney Junior), 2006. Produced by Disney Television Animation. 4 seasons, 125 episodes, 2006–2016.

[11] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[12] Wenyi Hong, Yean Cheng, Zhuoyi Yang, Weihan Wang, Lefan Wang, Xiaotao Gu, Shiyu Huang, Yuxiao Dong, and Jie Tang. MotionBench: Benchmarking and improving fine-grained video motion understanding for vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[13] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[14] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

[15] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. MVBench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22195–22206, 2024.

[16] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 5971–5984, 2024.

[17] Deborah L Linebarger and Dale Walker. Infants' and toddlers' television viewing and language outcomes. *American Behavioral Scientist*, 48(5):624–645, 2005. doi: 10.1177/0002764204271505.

[18] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2630–2640, 2019.

[19] Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. Teaching CLIP to count to ten. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3170–3180, 2023.

[20] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[21] Ziyao Shangguan, Chuhan Li, Yuxuan Ding, Yanan Zheng, Yilun Zhao, Tesca Fitzgerald, and Arman Cohan. TOMATO: Assessing visual temporal reasoning capabilities in multimodal foundation models. In *International Conference on Learning Representations (ICLR)*, 2025.

[22] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[23] Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, Ruochen Xu, and Tiancheng Zhao. VLM-R1: A stable and generalizable R1-style large vision-language model. *arXiv preprint arXiv:2504.07615*, 2025.

[24] Yueqi Song, Tianyue Ou, Yibo Kong, Zecheng Li, Graham Neubig, and Xiang Yue. VisualPuzzles: Decoupling multimodal reasoning evaluation from domain knowledge. *arXiv preprint arXiv:2504.10342*, 2025.

[25] Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. Coin: A large-scale dataset for comprehensive instructional video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1207–1216, 2019.

[26] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5238–5248, 2022.

[27] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9568–9578, 2024.

[28] Valerie Walsh Valdes, Chris Gifford, and Eric Weiner. Dora the explorer. Television series, 2000–2019. Nickelodeon Animation Studio.

[29] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihuan Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[30] Shengao Wang, Arjun Chandra, Aoming Liu, Venkatesh Saligrama, and Boqing Gong. BabyVLM: Data-efficient pretraining of VLMs inspired by infant learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1380–1390, 2025.

[31] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *European conference on computer vision*, pages 396–416. Springer, 2024.

[32] Zhiqiang Wang, Pengbin Feng, Yanbin Lin, Shuzhang Cai, Zongao Bian, Jinghua Yan, and Xingquan Zhu. CrowdVLM-R1: Expanding R1 ability to vision language model for crowd counting using fuzzy group relative policy reward. *arXiv preprint arXiv:2504.03724*, 2025.

[33] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786, 2021.

[34] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *ACM Multimedia*, 2017.

[35] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. CLEVRER: Collision events for video representation and reasoning. In *International Conference on Learning Representations*, 2020. URL `https://arxiv.org/abs/1910.01442`.

[36] Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, et al. Lmms-eval: Reality check on the evaluation of large multimodal models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 881–916, 2025.

[37] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Llava-video: Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024.

[38] Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, and Josef Sivic. Cross-task weakly supervised learning from instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3537–3545, 2019.

A Hyperparameter Sweeps

We conduct sequential hyperparameter sweeps to identify optimal GRPO training configurations. Each sweep fixes previously optimized hyperparameters while varying a single dimension.

Table 8: Sequential hyperparameter sweeps for GRPO training. Each sweep fixes previously optimized parameters. Bold indicates selected values used in main experiments.

Parameter	Values Tested	Selected	DoraVQA
KL Coefficient	{0.0, 0.001, **0.01**, 0.1}	**0.01**	{42.96, 40.71, **46.38**, 45.19}
Learning Rate	{1e-6, 1e-5, **1e-4**, 1e-3}	**1e-4**	{43.69, 45.19, **50.09**, 38.89}
Reward Scaling	{0.5, 1.0, **2.0**, 5.0}	**2.0**	{45.92, 45.19, **45.69**, 43.93}

Table 8 presents our hyperparameter optimization results. The KL coefficient controls the divergence penalty between the policy and reference model; we find 0.01 provides the best balance, with both higher (0.1) and lower (0.001) values degrading performance. For learning rate, 1e-4 achieves optimal performance (50.09%), while higher rates (1e-3) cause training instability (38.89%) and lower rates (1e-6) result in insufficient learning. Reward scaling at 2.0 provides marginal improvements over the default 1.0, amplifying the learning signal from correct answers without overwhelming the policy gradient.

Table 9 examines training duration and overfitting characteristics. Qwen2-VL-7B peaks at 150 steps (62.96%) before degrading sharply at 250 steps (56.0%), indicating overfitting to the narrow educational domain. Qwen3-VL-2B shows more robust training, peaking later at 250 steps (67.97%) with gradual performance decline. These results highlight the importance of early stopping when fine-tuning on domain-specific content to preserve generalization capabilities.

Table 9: Ablation study on training steps for both models. Performance peaks at between 150 and 250 steps then degrades due to overfitting on the narrow educational domain.

Model	50	100	150	200	250	300
Qwen2-VL-7B	57.49	60.29	**62.96**	60.98	56.00	60.84
Qwen3-VL-2B	63.67	66.67	65.91	65.54	**67.97**	66.60

Based on these sequential sweeps, we use KL coefficient = 0.01, learning rate = 1e-4, and reward scaling = 2.0 for all main experiments reported in this paper. Training duration is set at 150 steps for Qwen2-VL-7B and 250 steps for Qwen3-VL-2B to avoid overfitting while maximizing in-domain performance.

Reward-weight sensitivity. Table 10 varies the F1/Levenshtein weights on Qwen3-VL-8B. Every tested setting improves both transfer benchmarks over the pretrained baseline (71.40/62.10), but 3:7 performs best among these settings. This supports giving the edit-distance term greater weight while retaining F1's tolerance for surface variation; it does not establish a universal optimum.

Table 10: Reward-weight sweep on Qwen3-VL-8B + GRPO. Scores are accuracy (%).

$\alpha : \beta$	Video-MME	NExT-QA
1:9	74.78	72.25
3:7	**80.00**	**81.80**
5:5	74.85	78.00
7:3	77.61	78.33
9:1	75.15	77.17

B Pause Visual Hint Analysis

A central hypothesis of this work is that the *pause* segment in the *context–question–pause–answer* structure functions as an implicit reasoning trace: it visually isolates the evidence required to answer the question, thereby grounding language in perception. To test whether this signal is recoverable by an off-the-shelf vision-language model, we design a controlled prompting experiment that makes the show's blue-arrow hint explicitly available to the model, and measure how much of the resulting gain requires GRPO fine-tuning versus prompt engineering alone. To enable this analysis and support future spatial-grounding work, we manually annotated 702 bounding boxes around the blue directional arrows that appear during *Dora the Explorer* pause segments, released alongside the SoSVQA pipeline.

Setup. We evaluate four conditions on a held-out subset of DoraVQA examples that contain a visible blue directional arrow during the pause segment. The conditions form an intervention-strength ladder over the spatial signal available at inference:

- **No hint.** Standard inference; the model is not told the arrow exists.
- **Self-localized.** The model emits a bounding box for the arrow in a first pass and conditions on its own (noisy) prediction in a second pass.
- **Prompted hint.** The model receives a textual instruction: *"Locate the blue arrow in the image. Use its bounding-box coordinates to identify the object or region it is pointing at, and use that information to answer the question."*—no coordinates are produced or supplied externally.
- **Oracle box.** The ground-truth bounding box of the arrow is supplied directly in the prompt.

We additionally evaluate each condition with our **+ GRPO** fine-tuned model. All conditions use identical visual frames and question text; the only variables are the system prompt and whether the model weights have been fine-tuned.

Qualitative findings. Figure 4 illustrates two representative examples. In *Example 1* ("Who is the strongest one in the team?"), the baseline hallucinates an incorrect character. The oracle box redirects visual attention to the correct spatial region, but the model describes the referent only at the level of a perceptual property ("the blue animal"), failing to resolve the arrow to a named entity. Only GRPO fine-tuning enables the full inference chain—from arrow location, to object identity, to the correct named answer (Benny the bull). In *Example 2* ("What makes you sneeze?"), the hint alone is sufficient: the arrow unambiguously isolates a single salient object (the pepper shaker), requiring no entity-level knowledge to name correctly. Both oracle conditions succeed here, while the no-hint condition fails entirely.

Interpretation. These examples reveal a consistent two-stage failure mode in models without GRPO. First, without the hint, the model does not attend to the pause arrow at all, defaulting to plausible but incorrect priors from its pre-training distribution. Second, even when the hint redirects attention, the model can localize a region without being able to resolve it to a named entity—a gap that GRPO training on pedagogically structured content bridges by repeatedly supervising the full reasoning chain from visual evidence to grounded answer. This suggests that the pause arrow encodes two distinct learning signals: a *spatial attention* signal (where to look) that is recoverable through prompting, and an *entity grounding* signal (what it is called) that requires the kind of structured supervision our dataset provides. The quantitative results for this ablation are reported in Table 7.

Scaling behavior. To test whether scale alone recovers the pause signal, we evaluate the same conditions on larger untuned models (Table 11). Qwen2.5-VL-72B with prompted hint (60.66%) trails the GRPO-finetuned 8B model (67.61%), confirming that scale partially recovers spatial attention but does not replace structured supervision for entity grounding. Qwen2.5-VL-72B does achieve 11.87% mean IoU for arrow localization vs. $<3\%$ for the 8B models, indicating that spatial localization benefits more directly from scale than entity resolution does.

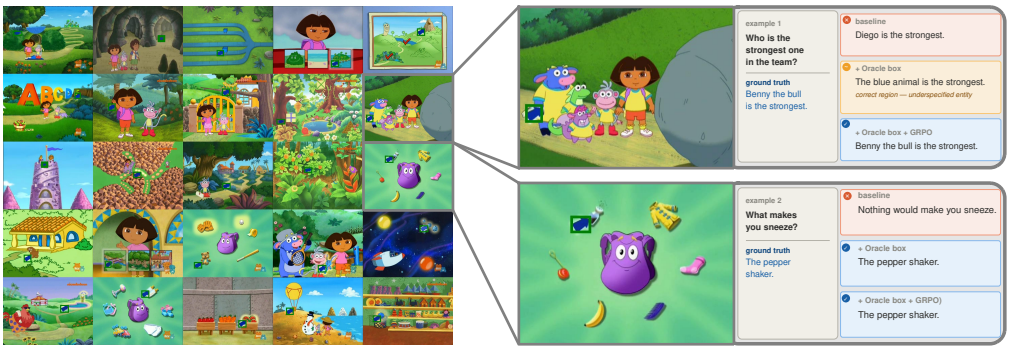


Figure 4: **Effect of pause-based visual hints on spatial grounding.** We evaluate three conditions on two examples drawn from *Dora the Explorer*'s pedagogical pause segments, where a blue arrow highlights the visually salient object relevant to the question. **(No hint)** receives only the raw frame and question with no additional guidance. **(+ Oracle box)** additionally supplies the ground-truth bounding box of the arrow's target region as part of the prompt. **(+ Oracle box + GRPO)** combines the same oracle box with our GRPO-finetuned model. In *Example 1*, the baseline hallucinates an incorrect character (Diego); the oracle box redirects attention to the correct spatial region but the untuned model describes the referent only at the level of a perceptual property ("the blue animal"), failing to resolve it to a named entity. Only GRPO fine-tuning enables the model to bridge from spatial location to the correct named answer (Benny the bull). In *Example 2*, both conditions succeed, confirming that an unambiguous spatial hint is sufficient for object-level queries that require no entity-level knowledge. Together, these results demonstrate that the show's pause arrow functions as an *implicit reasoning trace*: it constrains the visual search space, but generalizing from spatial localization to grounded entity recognition requires the structured supervision provided by GRPO training on pedagogical content.

Table 11: Pause-based visual hint results across model scales (blue-arrow subset). Same setup and conditions as Table 7. Qwen3-VL-8B + GRPO (8B) surpasses both Qwen3-VL-30B and Qwen2-VL-72B on overall accuracy, while Qwen2.5-VL-72B is the only untuned model with meaningful localization (11.87% mIoU). The *self-localized* condition degrades Qwen3-VL-30B sharply (35.19 vs. 55.52 overall), revealing that the coordinate-emission pathway is unstable at this scale—whereas the *prompted hint*, which carries no coordinates, is robust across all model sizes. **Bold**: best overall per column. Underline: best per model.

Model	Condition	Overall	Spatial	Nav.	Know.	Loc. (mIoU)
Qwen3-VL-8B	No hint	52.24	58.17	35.38	56.10	–
	Self-localized	53.61	55.83	39.52	64.86	0.76
	Prompted hint	55.94	64.59	31.54	57.14	–
	Oracle box	55.25	61.24	45.60	46.15	–
Qwen3-VL-8B + GRPO	No hint	63.19	70.48	34.62	52.38	–
	Self-localized	66.98	**80.00**	36.92	**69.05**	2.57
	Prompted hint	**67.61**	77.14	**46.92**	52.38	–
	Oracle box	**67.61**	77.62	43.85	64.29	–
Qwen3-VL-30B	No hint	55.52	63.24	25.81	48.78	–
	Self-localized	35.19	39.80	19.49	39.47	0.00
	Prompted hint	54.37	63.05	28.00	51.28	–
	Oracle box	52.69	59.90	30.17	61.11	–
Qwen2-VL-72B	No hint	52.85	57.62	31.78	52.38	–
	Self-localized	56.24	63.81	36.15	50.00	2.03
	Prompted hint	55.45	64.29	33.85	64.29	–
	Oracle box	55.45	61.90	33.85	59.52	–
Qwen2.5-VL-72B	No hint	56.56	61.90	40.00	50.00	–
	Self-localized	54.19	56.19	34.62	61.90	**11.87**
	Prompted hint	60.66	66.19	40.77	59.52	–
	Oracle box	58.77	65.71	42.31	64.29	–

Table 12: Arrow localization accuracy across model scales (*self-localized* condition only, where coordinates are externalized). IoU@k reports the percentage of predictions exceeding the given threshold. The GRPO-finetuned 8B model (2.57% mIoU) outperforms both the 30B and 72B untuned models, demonstrating that structured supervision recovers spatial localization more effectively than scaling alone.

Model	IoU@10	IoU@25	IoU@50	IoU@75	mIoU
Qwen3-VL-8B	0.00	0.00	<1	<1	0.76
Qwen3-VL-8B + GRPO	0.00	0.00	<1	<1	2.57
Qwen3-VL-30B	0.00	0.00	0.00	0.00	0.00
Qwen2-VL-72B	6.96	0.63	0.00	0.00	2.03
Qwen2.5-VL-72B	**35.51**	**20.22**	**5.10**	**0.80**	**11.87**

C Additional Qualitative Results

Figure 5 presents two challenging examples from SoSVQA that illustrate distinct failure modes of baseline models. The spatial location example (left) demonstrates partial occlusion: Swiper the fox is only partially visible behind Dora, requiring the model to recognize an object from incomplete visual information. Without GRPO fine-tuning, Qwen2-VL-2B fails to detect the occluded character entirely, while the GRPO-finetuned variant correctly identifies his location consistent with the spatial perception gains reported in Table 1. The deixis resolution example (right) illustrates a subtler failure mode. The question "Mickey, what's that?" cannot be resolved by identifying the most visually salient object in the frame; it requires grounding the referent *what* in the episode's narrative context. Baseline models default to describing visually dominant elements (the colorful bus) or arbitrarily salient objects (a tree stump), answering "what is visually prominent?" rather than "what is Mickey referring to?". Our GRPO-finetuned models correctly resolve the reference to the drum by grounding the question in the episodic context; the same entity grounding signal that the *context-question-pause-answer* structure supervises directly (Section B). Together, these examples demonstrate that GRPO fine-tuning on pedagogically structured content improves both spatial perception under occlusion and narrative-grounded reference resolution. **Additional interactive qualitative results.** We provide an interactive HTML page in the supplementary materials with further visual examples across all task types. Please open `supplementary/index.html` in a web browser to explore the full set of qualitative results beyond the examples shown in this paper.

D Prompt Templates

To support full reproducibility, this section documents the exact message structures used during GRPO training and multiple-choice evaluation. Visual frames are uniformly sampled from the video segment associated with each question; placeholders of the form <...> are filled in per-example at runtime.

D.1 Frame Sampling

For every training and evaluation example, we uniformly sample N frames from the video segment $[t_{\text{start}}, t_{\text{end}}]$ associated with the question, where t_{start} is the beginning of the *context* segment and t_{end} is the end of the *pause* segment. Sampled frames are denoted <*frame_1*>, . . . , <*frame_n*>, and are passed in temporal order. We use N=4 unless otherwise noted; the prompt structure is independent of N.

D.2 GRPO Training Prompt

During GRPO training, the model receives the sampled frames together with the episode transcript (with <*pause*> tokens inserted at scene boundaries to mark the pedagogical pause), the question, and an open-ended `Answer:` cue. Free-form generations are scored against the ground-truth answer a^* using the reward defined in Eq. 1.

GRPO Training Message (Open-Ended Generation)

```
messages = [
    {
        "role": "user",
        "content": [
            {"type": "image", "image": <frame_1>},
            {"type": "image", "image": <frame_2>},
            # ... uniformly sampled frames in temporal order ...
            {"type": "image", "image": <frame_n>},
            {
                "type": "text",
                "text": (
                    "You are a helpful visual reasoning assistant for kids.\n"
                    "Think step by step, then give a final concise answer.\n"
                    "Context: <transcript_with_pause_tokens>\n"
                    "Question: <question_text>\n"
                    "Answer:"
                ),
            },
        ],
    },
]
```

D.3 Multiple-Choice Evaluation Prompt

All cross-benchmark evaluations (Video-MME, NExT-QA, MVBench, MotionBench, TVBench) and the in-domain SoSVQA evaluation use the multiple-choice format below. The model is constrained to emit a single digit selecting one of four choices, which we parse directly. As noted in Section 3.2, this deliberate format mismatch with training isolates transferable reasoning from answer-format memorization.

Multiple-Choice Evaluation Message (Single-Digit Output)

```
messages = [
    {
        "role": "user",
        "content": [
            {"type": "image", "image": <frame_1>},
            {"type": "image", "image": <frame_2>},
            # ... uniformly sampled frames in temporal order ...
            {"type": "image", "image": <frame_n>},
            {
                "type": "text",
                "text": (
                    "You are a helpful visual reasoning assistant for kids.\n"
                    "Think step by step and choose one correct choice "
                    "from 0, 1, 2 or 3.\n"
                    "Only return one single digit as the best answer.\n"
                    "Context: <transcript_with_pause_tokens>\n"
                    "Question: <question_text>\n"
                    "Choices: {\"0\": \"choice_A\", \"1\": \"choice_B\", "
                    "\"2\": \"choice_C\", \"3\": \"choice_D\"}"
                ),
            },
        ],
    },
]
```

Component	Training (GRPO)	Evaluation (MCQ)
Frames	N uniformly sampled, temporal order	N uniformly sampled, temporal order
Context	Transcript with <*pause*> tokens	Transcript with <*pause*> tokens
Question	Free-form question text	Free-form question text
Choices	✗ (none)	✓ JSON dict, keys 0–3
Output target	Open-ended answer string	Single digit $\in \{0, 1, 2, 3\}$
Scoring	Reward (Eq. 1) vs. a^*	Exact-match top-1 accuracy

Table 13: Differences between training and evaluation prompts. The training prompt is open-ended; the evaluation prompt adds explicit multiple-choice formatting and constrains the output to a single digit.

D.4 Summary of Prompt Components

E Data Access and Licensing

SoSVQA does not redistribute any video frames, audio, or raw episode files. All visual and audio content was collected from publicly available YouTube uploads and remains hosted by the original rights holders. Users are required to independently obtain access to the corresponding episodes under YouTube's existing terms of service. We release episode identifiers, temporal timestamps, question-answer annotations, choices, transcript span indices, manually annotated bounding box for the pause-segment visual cues along with the code to align these annotations. As the source material consists of publicly broadcast children's educational programming, the datasets do not introduce new privacy risks; nevertheless, SoSVQA is intended strictly for research and evaluation purposes, and we discourage any use outside this scope.

F MCQ Generation and Manual Auditing

To convert the open-ended QA pairs extracted from the *context-question-pause-answer* structure into a multiple-choice evaluation format, we use Google Gemini 2.5 Flash to (i) simplify the gold answer into a concise, MCQ-appropriate phrase and (ii) generate three plausible but incorrect distractors of similar style and length. The text-only prompt operates on the question, gold answer, and surrounding transcript context (truncated to the most recent 1,000 characters when longer), and returns a structured JSON object with the simplified answer and distractors. The exact prompt template is shown below.

Gemini MCQ Generation Prompt (Text-Only)

```
Create a multiple-choice question with <num_distractors> wrong answer options.

VIDEO CONTEXT:
<transcript_context>

QUESTION: <question_text>
ORIGINAL CORRECT ANSWER: <gold_answer>

TASK:
1. First, simplify the original correct answer to be concise and
   MCQ-appropriate (similar length to distractors, clear and direct)
2. Then generate <num_distractors> wrong but plausible answers that match
   the answer style and length

REQUIREMENTS:
- Simplified answer: Short, clear, direct (1-2 sentences max,
  ideally 1 phrase)
- Distractors: Each must be incorrect, believable, simple words for
  ages 3-7, unique, and similar length to the simplified answer

OUTPUT FORMAT - Copy this structure exactly and fill in all fields:
```

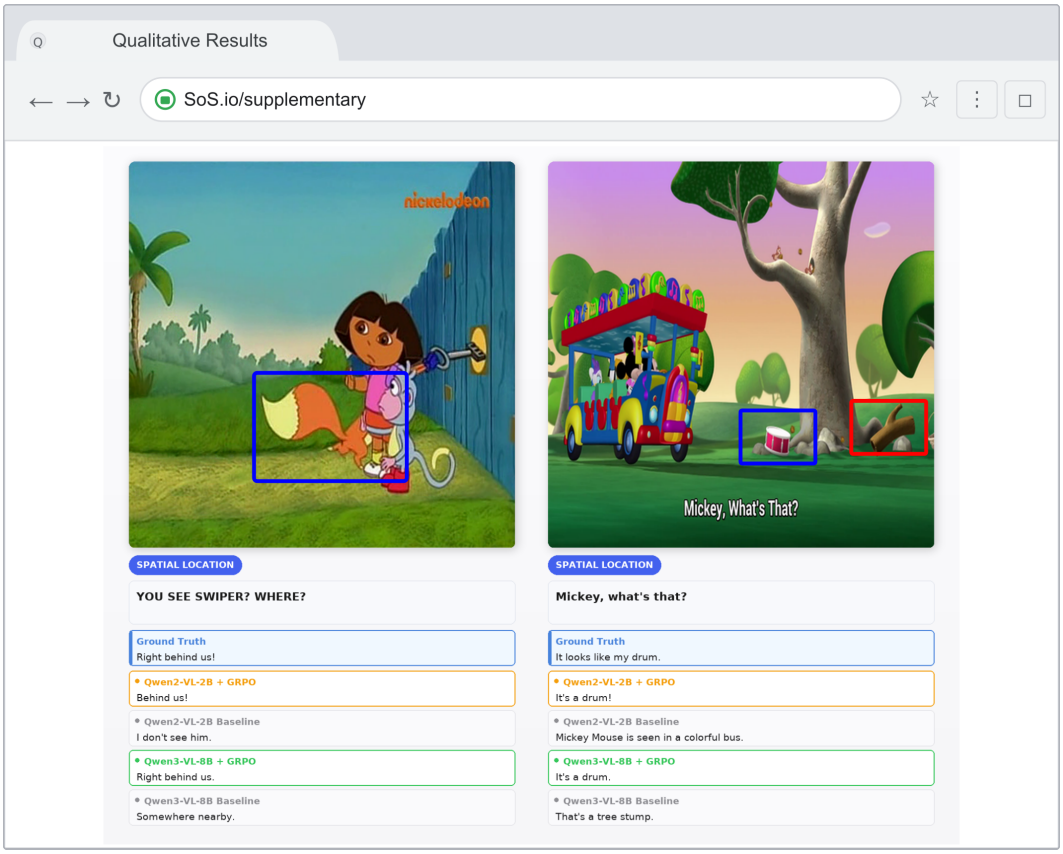


Figure 5: **Additional challenging examples from SoSVQA.** Left: Spatial location task where Swiper the fox is partially occluded behind a blue wall, requiring the model to recognize objects from incomplete visual information (blue box). Right: A deixis resolution task where the question "Mickey, what's that?" requires understanding *what* Mickey is referring to within the episodic narrative context. Baseline models either describe the dominant visual element (the colorful bus) or latch onto an unrelated object—the tree stump (red box)—revealing that they answer "what is visually prominent?" rather than "what is Mickey referring to?". Our GRPO-finetuned models correctly resolve the reference to the drum (blue box) by grounding the question in the episode's narrative context. **Additional interactive qualitative results.** We provide an interactive HTML page in the supplementary materials with further visual examples across all task types. Please open `supplementary/index.html` in a web browser to explore the full set of qualitative results beyond the examples shown in this paper.

```
{
    "simplified_answer": "concise version of the correct answer",
    "distractor_1": "your first wrong answer",
    "distractor_2": "your second wrong answer",
    "distractor_3": "your third wrong answer"
}

IMPORTANT:
- Simplify the answer to be concise (like the distractors will be)
- You must provide simplified_answer, distractor_1, distractor_2, AND
  distractor_3. All fields are required.
```

Generation uses temperature 0.5, top-$p = 0.95$, and a maximum of $2{,}048$ output tokens. Each candidate MCQ is post-processed to enforce case-insensitive uniqueness between the simplified answer and distractors, and to discard any item that does not yield at least three unique distractors after deduplication.

Manual auditing of deictic answers. Although Gemini reliably produces fluent distractors, a recurring failure mode arises from the pedagogical structure itself: many answers in the source material are *deictic*—phrases such as "there it is," "here it is," or "right there" that derive their meaning from the on-screen visual hint (e.g., a directional arrow or character gesture) rather than from any named entity. When such a phrase becomes the gold MCQ option alongside three named-entity distractors generated from the transcript, the resulting item is trivially solvable by surface form alone: the deictic phrase is the only choice that does not name a specific entity, leaking the correct answer without requiring any visual reasoning.

To eliminate this shortcut, we manually audited every MCQ whose simplified answer or original answer contained a deictic marker (*there it is*, *here it is*, *right there*, and minor morphological variants). The audit identified **106 affected items in DoraVQA and 31 in ClubHVQA**. For each affected item, we manually rewrote the four choices so that all options share a consistent surface form—typically a yes/no acknowledgement paired with a directional or perceptual claim—forcing the model to ground the answer in the visual scene rather than in lexical type. Figure 6 shows a representative before/after example.

Concretely, the question *"Now, where's the quiet forest? Do you see the quiet forest?"* originally paired the deictic gold answer *"There it is!"* against three named-location distractors (*Little bird?*, *Yellow Valley?*, *Big mountain?*), so the answer could be selected purely by recognizing which option is not a named location. After auditing, the four choices become parallel yes/no–directional statements (*Yes, it's on the left.*, *No, I can't see it.*, *Yes, it's on the right.*, *I see flowers.*), which can only be resolved by attending to the on-screen arrow during the pause segment.

This audit affects $106/5{,}348 = 1.98\%$ of DoraVQA and $31/4{,}652 = 0.67\%$ of ClubHVQA. The lower rate in ClubHVQA is consistent with its more visually-driven, object-selection-dominated structure (78% object selection; Figure 1), in which gold answers are typically named objects rather than deictic references. All audited items are released as part of the SoSVQA annotations, with both the original and revised choice sets preserved for transparency.

G Limitations

While SoSVQA demonstrates that pedagogical structure enables stronger spatial reasoning abilities for VLMs despite limited data scale, the approach has several limitations. First, the dataset depends on the availability and quality of publicly hosted videos and corresponding transcripts; oftentimes the alignment between SRT files and episode timestamps is imperfect and requires human input. Misaligned or noisy subtitles can introduce extraction errors, especially in fast-paced scenes or episodes with inconsistent caption timing. Second, SoSVQA currently draws from only two children's educational programs–Dora the Explorer and Mickey Mouse Clubhouse. Although these shows share a strong context-question-pause-answer structure, they represent a narrow slice of pedagogical video, and broader generalization would require incorporating additional sources with diverse reasoning styles and visual environments. Third, despite improvements from GRPO fine-tuning, counting performance remains a persistent challenge: models still struggle with fine-grained enumeration, and

Question: *Now, where's the quiet forest? Do you see the quiet forest?*
Gold answer: *Yes– there it is* $\rightarrow$ *There it is!*



Pause segment: Boots is pointing to the quiet forest on the right.

Before audit : shortcut leak

- Little bird?
- Yellow Valley?
- **There it is! ✓**
- Big mountain?

The deictic gold answer is the only non-entity option—selectable by surface form alone.

After audit : visually grounded

- Yes, it's on the left.
- No, I can't see it.
- **Yes, it's on the right. ✓**
- I see flowers.

All four options share a parallel yes/no–directional surface form, requiring the model to ground the answer in the on-screen arrow.

Figure 6: Manual audit of deictic gold answers in SoSVQA. The original choice set leaks the answer through surface form; "There it is!" is the only non-named-entity option and can be selected without attending to the visual scene. The audited choice set forces visual grounding by making all four options parallel directional statements. We applied this audit to all 106 deictic items in DoraVQA and 31 in ClubHVQA.

the gains observed in counting tasks are modest relative to other reasoning categories. Fourth, due to limited hardware availability, we evaluate only a small set of open-source VLM families and parameter scales. Larger models and alternative architectures may benefit more from structured supervision, but we were unable to explore these directions. Copyrighted audiovisual content cannot be redistributed with the annotations, so reproducibility depends on users obtaining independent access. Familiar shows may occur in pretraining data; the recall and visual counterfactual probes reduce, but do not eliminate, contamination concerns. Transfer to new pedagogical sources remains untested, and CLEVRER changes are mixed. Together, these limitations highlight opportunities for expanding the dataset, improving alignment quality, and scaling evaluations to a wider range of model families and pedagogical sources.

H Compute Resource

For the Qwen2-7B and Qwen3-8B models, all GRPO fine-tuning runs were trained for 450 epochs on a single NVIDIA H200 GPU with an 8-core CPU and 64 GB of system memory. Each model used a combined batch size of 8, gradient accumulation of 4, and 8 sampled generations per prompt. Optimization followed standard GRPO settings, with a learning rate of 1e-4, KL-regularization coefficient β=0.01, and reward weight of 2.0, informed by our parameter sweep results. For the 2B variants, training was performed on a single NVIDIA A100 GPU, while all other hyperparameters and training configurations remained unchanged. Each individual fine-tuning run required approximately 7-8 GPU hours on the H200 or A100, depending on the model size and checkpoint-resume behavior. Across all models, ablations, and joint-training configurations, the total compute used for the experiments reported in the paper amounts to roughly 120-150 GPU-hours. The full research project required additional compute beyond the final reported experiments, including preliminary GRPO reward weight sweeps, early alignment-quality diagnostics, and several failed or discarded runs; these exploratory stages added an estimated 40-60 GPU-hours.

Inference was conducted using the default settings provided by LMMS-Eval [36] and VLLM [14], running across a heterogeneous pool of GPUs–including NVIDIA V100, A100, A5000, A6000, H200, and L40S–selected based on model size and input sequence length. Storage requirements were modest: each run required approximately 50-80 GB of temporary workspace for checkpoints,

logs, and intermediate outputs. All experiments were conducted on an private high-performance computing (HPC) cluster equipped with GPU nodes accessible through a shared scheduling system.

I Safeguards

In our work, we take several safeguards to ensure responsible use and to minimize risks of misuse or unintended deployment. First, we do not redistribute any copyrighted video frames, audio, or raw episode content. All visual and audio material remains hosted by the original rights holders on YouTube, and users must independently obtain access under YouTube's existing terms of service. We release only episode identifiers, timestamps, transcript spans, question–answer annotations, and manually annotated bounding boxes, which pose minimal risk and contain no personally identifiable information. Second, because the source material consists of publicly broadcast children's educational programming, the dataset does not introduce new privacy risks; nonetheless, we explicitly state that SoSVQA is intended strictly for research and evaluation, and we discourage any use outside this scope.

To further reduce misuse risk, we will provide clear documentation describing the dataset's structure, intended purpose, and limitations, and we avoid releasing any large-scale pre-trained or instruction-tuned models that could be repurposed for harmful applications. Our fine-tuned models are trained only on pedagogically structured QA pairs and do not include any safety-critical capabilities. Finally, we will include usage guidelines in the repository emphasizing that the dataset should not be used for commercial deployment, content generation, or applications involving children, and we provide scripts that require users to supply their own video access, preventing accidental redistribution of copyrighted or sensitive material.

Together, these safeguards ensure that the released assets are responsibly scoped, legally compliant, and suitable for research use without enabling high-risk downstream applications.

J Additional Backbone Evaluations

InternVL3-8B and Qwen3.5-4B were fine-tuned with the same SoSVQA GRPO recipe and evaluated on both source subsets and the two primary transfer benchmarks. Table 14 collects the available results. Table 1 includes their overall SoSVQA scores, with dashes for unreported category breakdowns. We keep their transfer results here rather than in Table 2, since the remaining three transfer benchmarks were not evaluated for these runs.

Table 14: Additional-backbone results (accuracy, %). Each GRPO row uses the identical training recipe and is compared with its own baseline.

Model	DoraVQA	ClubHVQA	Video-MME	NExT-QA
InternVL3-8B	55.21	50.36	68.90	82.57
InternVL3-8B + GRPO	**61.73**	**57.61**	**75.68**	**84.23**
Qwen3.5-4B	46.17	39.06	63.59	72.70
Qwen3.5-4B + GRPO	**52.29**	**50.96**	**65.26**	**79.17**

Both backbones improve on every reported dataset. InternVL3-8B gains 6.52/7.25 points on DoraVQA/ClubHVQA and 6.78/1.66 on Video-MME/NExT-QA; Qwen3.5-4B gains 6.12/11.90 and 1.67/6.47, respectively. These comparisons broaden the evidence beyond the original Qwen backbones without implying a complete evaluation on all benchmarks.

K Additional Transfer and Reproducibility Results

Variation across GRPO runs. We repeat Qwen3-VL-8B GRPO training with three additional random seeds and otherwise identical settings. Table 15 reports each run. The standard deviation is the sample standard deviation across the four runs, $\sqrt{\sum_i (x_i - \bar{x})^2/(4-1)}$; it measures training-seed variability, not a confidence interval over evaluation questions.

Table 15: Qwen3-VL-8B GRPO run-to-run variation (accuracy, %).

Run	Video-MME	NExT-QA
Reported run	80.00	81.80
Additional seed 1	79.98	80.00
Additional seed 2	80.09	82.00
Additional seed 3	81.14	81.80
Mean $\pm$ sample std.	80.30 ± 0.56	81.40 ± 0.94

Far-domain transfer: CLEVRER. On CLEVRER [35], we evaluate 500 descriptive and 500 multiple-choice questions from the validation split, sampled with seed 0, using 16 frames. Table 16 reports accuracy as fractions. Changes are small and mixed: Qwen3-VL-8B improves slightly on all metrics, while the 2B model declines on descriptive and per-option accuracy and the 7B model declines slightly on descriptive accuracy. These results bound the transfer claim; they do not support uniform improvement on synthetic physical reasoning. Figure 7 illustrates the task format.

Table 16: CLEVRER validation probe. Paired cells are base / + GRPO; the frontier row is zero-shot. All use 16 frames, 500 descriptive and 500 MC questions (1,676 MC options); MC metrics score options or complete questions.

Model	Descriptive	MC (per-option)	MC (per-question)
Qwen2-VL-2B	**.746** / .736	**.633** / .624	.240 / **.254**
Qwen2-VL-7B	**.862** / .856	**.645** / .658	.236 / **.252**
Qwen3-VL-8B	**.800** / .804	**.751** / .762	.416 / **.424**
GPT-5-mini (zero-shot)	.488	.554	.188

Extraction-quality evidence. All MCQ items undergo manual auditing, including correction of 137 deictic-answer items (Appendix F); 702 pause-cue bounding boxes are manually annotated (Appendix B). Timestamp/subtitle alignment remains a source of noise (Appendix G).

L Additional Frontier and Abstract-Reasoning Probes

The frontier-model breakdowns are included in the separate subset block of Table 1: seed 0, 534 requested questions per show, four frames, and zero-shot MCQ prompting. Gemini-3.6-Flash has one unparseable DoraVQA response (533 scored); the other runs score all 534. Category values use the scored denominator. The object-selection/counting column pools the two categories by their question counts, and the knowledge column uses the factual-recall category. These subset results should not be treated as full-test head-to-head comparisons.

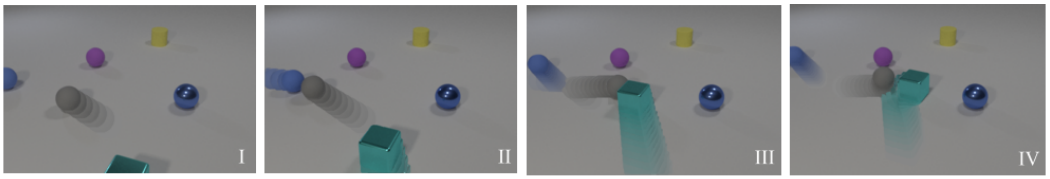
Static visual reasoning. VisualPuzzles [24] uses the full test split ($n = 1168$); SMART-101 [5] uses a seeded subset ($n = 606$). Both evaluate single images after video-based training; examples appear in Figures 8 and 9. Table 17 shows improvements of 3.01–3.99 points on VisualPuzzles. SMART-101 is nearly flat for the two smaller backbones, while Qwen3-VL-8B gains 5.54 points. These are descriptive single-run comparisons, without a statistical-significance claim. SMART-840 results are not included because access was still pending during the rebuttal.

Table 17: Static-puzzle accuracy (%). Each cell reports base / + GRPO.

Model	VisualPuzzles	SMART-101
Qwen2-VL-2B	31.30 / **34.31**	27.72 / **27.73**
Qwen2-VL-7B	30.20 / **33.28**	**27.74** / 27.23
Qwen3-VL-8B	34.42 / **38.41**	36.00 / **41.54**

M Visual Examples of the Additional Evaluation Datasets

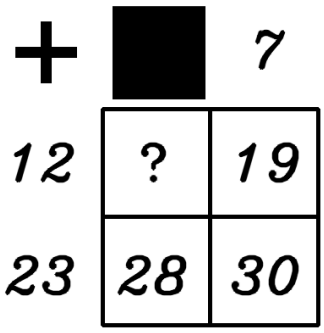
Figures 7–9 illustrate the visual inputs and question formats of the additional transfer benchmarks. They reproduce examples from the datasets' official project pages, rather than model predictions or a selection of successful evaluations. The examples are illustrative and are not asserted to belong to our seeded evaluation subsets.



Descriptive example
Count the moving spheres. The source example's answer is two.

Temporal and causal examples
Explain a collision, predict the next collision, or reason about events after removing an object.

Figure 7: **CLEVRER: physical-event reasoning.** Rendered objects undergo collisions, supporting descriptive, explanatory, predictive, and counterfactual questions. The supplied evaluation reports descriptive accuracy and MC scores for individual options and complete questions. Frames reproduced from the CLEVRER project page [35], cropped from its teaser; task descriptions are paraphrased from the source examples.



SMART-101: arithmetic puzzle

The image encodes a small addition table. The task asks for the number hidden by the question mark, combining the row and column relationships.

Options: A: 18; B: 21; C: 17; D: 20; E: 14.

This example illustrates reasoning over an image with embedded mathematical symbols, rather than a video or transcript.

Figure 8: **SMART-101 example.** Image reproduced from the SMART-101 project page [5] (puzzle 10). The accompanying task description is paraphrased. The dataset is distributed under CC BY-SA 4.0; source puzzle images are credited by the dataset authors to Math Kangaroo USA. No model output is shown.

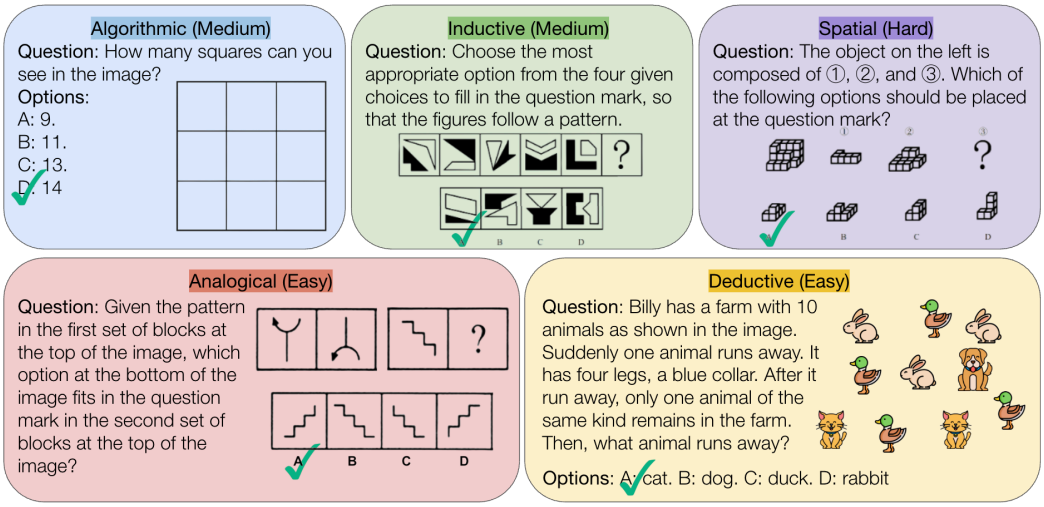


Figure 9: **VisualPuzzles: abstract visual reasoning.** Official examples span algorithmic, inductive, spatial, analogical, and deductive reasoning. The green checks belong to the source illustration and indicate its annotated answers; they are not predictions from our models. Figure reproduced from the VisualPuzzles project page [24].

The three datasets vary the input and reasoning demands beyond SoSVQA: CLEVRER supplies synthetic physical events, SMART-101 supplies mathematical image puzzles, and VisualPuzzles supplies abstract visual tasks. The quantitative comparisons remain in Tables 16 and 17; these illustrations explain the task formats rather than establish additional performance claims.

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: We believe that the abstract and introduction accurately restate the paper's core contributions, assumptions, and empirical scope without overstating claims or introducing elements that are not supported by the method or experiments.

   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: We provide detailed discussion regarding the weakness and limitations of our work in Appendix Sec. G.

   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 any formal theorems, proofs, analytical bounds, or other theoretical constructs; its contributions are entirely methodological (the SoSVQA pipeline), empirical (GRPO fine-tuning), and dataset-driven, so the category of "theoretical results" does not apply to this work.

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 fully specifies the SoSVQA extraction pipeline, dataset structure, training setup (including GRPO reward formulation, hyperparameters, prompt formats, and sampling procedures), and evaluation protocols, providing all methodological details needed 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: We will provides open access to the SoSVQA annotations, episode identifiers, timestamps, transcript spans, and the full extraction and training pipeline, along with clear instructions for aligning these annotations with publicly hosted video sources, which together are sufficient to faithfully reproduce all main experimental results.

   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: We specify all training and test details needed for understanding and reproducing the results, including dataset splits for DoraVQA and ClubHouseVQA, frame-sampling procedures, GRPO training hyperparameters (learning rate, KL coefficient, reward weights, group size), optimizer setup, prompt formats, evaluation protocols, and the exact benchmarks used, all of which are documented in the Method, Experiments, and Appendix sections.

   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: [No]

Justification: We report mean and sample standard deviation across four Qwen3-VL-8B GRPO runs on Video-MME and NExT-QA (Appendix K), capturing training-seed variability. Other experiments remain single-run results, so uncertainty coverage is incomplete.

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: We provide detailed compute resources usage report in Appendix Sec. H.

   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: We confirm that the research conducted in the paper conform with the NeurIPS Code of Ethics in every respect.

   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: [N/A]

    Justification: Our paper introduces a dataset and methodology derived from publicly available educational video content and evaluates model performance on spatial reasoning tasks, without influencing users, or introducing applications that would create direct societal benefits or risks. The work is limited to dataset construction and empirical evaluation on existing benchmarks, and does not involve user-facing systems, sensitive domains, or real-world deployment scenarios that would generate direct societal impact. As such, the work has no immediate societal impact.

    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: [Yes]

    Justification: Although our work does not pose high risk for misuse, we implemented several safeguards to ensure responsible release of data and models. Detailed information can be found in Appendix Sec. I.

    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: In our work, we explicitly credits the original creators of all external assets (e.g., Dora the Explorer and Mickey Mouse Clubhouse), does not redistribute any copyrighted video or audio, and clearly states that all source material remains hosted by the original rights holders under YouTube's terms of service. The paper releases only episode identifiers, timestamps, annotations, transcript spans, and alignment code, all of which are permissible to share, and it explicitly instructs users to obtain the underlying episodes independently and to use the dataset strictly for research purposes, ensuring full compliance with licensing and terms of use. For more detailed information, please refer to Appendix Sec. E.

    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: In our work, we introduce SoSVQA, which includes the following new assets: annotations, episode identifiers, timestamps, question–answer pairs, multiple-choice options, transcript span indices, and manually annotated bounding boxes for pause-segment visual cues. These assets are fully documented in the paper and supplemental material, and are released together with alignment code and usage instructions. The paper explains how each asset is constructed, how it should be used, and how it aligns with publicly available video sources, ensuring that all new resources are accompanied by clear and sufficient documentation.

    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: Our work does not involve crowd-sourcing nor researching with human subjects.

    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: Our work does not involve crowd-sourcing nor researching with human subjects.

    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: [Yes]

    Justification: The paper provides a complete and transparent description of how large language models (specifically Qwen-based model family) are used as a central and non-standard component of the method: it details the our supervision pipeline, the GRPO fine-tuning setup, the reward formulation, prompt formats, and the role of the VLM in learning structured spatial reasoning from educational video. This usage is not incidental but core to the contribution, and the paper documents each step necessary to understand how the LLM is integrated into data extraction, training, and evaluation.

    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.