Perception Before Reasoning: Dynamic Latent Reasoning for Video Understanding and Question Answering

Anonymous ACL submission

Abstract

Video question answering requires models to ground language queries in visual evidence and, when necessary, reason over that evidence across time. Existing methods typically rely on long textual chain-of-thought rationales, even though many questions can be answered as soon as the relevant object, action, or frame is localized. We propose Dynamic Latent Reasoning (**DyLaR**), which first grounds a question in latent visual evidence and then adaptively decides whether additional latent reasoning is needed. DyLaR produces a short block of *perception latents* that encode query-relevant visual evidence, then conditionally allocates *reasoning latents* only when the evidence must be composed before the final answer. To train this response format, we use localized evidence supervision for perception latents and rationale-based latent supervision for reasoning latents, followed by reinforcement learning with verifiable rewards. Across seven video benchmarks, DyLaR achieves a strong accuracy–efficiency trade-off. On Qwen3-VL-4B-Instruct, DyLaR improves average accuracy over Qwen3-VL-4B-Thinking from 55.1 to 56.7 while reducing generated response length from 2,431 to 16.6 tokens per query. Ablations further show that both grounded perception latents and rationale-supervised reasoning latents are important for video question answering accuracy. We will publicly release code and models.

1 Introduction

Multimodal large language models (MLLMs) have recently shown strong capabilities in video understanding and reasoning (Maaz et al., 2024; Lin et al., 2024; Zhang et al., 2025a). Recent advances further improve video reasoning with chain-of-thought supervision, reinforcement learning, and explicit evidence grounding (Feng et al., 2025; Wang et al., 2025a; Meng et al., 2025). However, the amount of additional reasoning required varies substantially across questions. Video question answering requires models to identify the visual evidence relevant to a language query and use that evidence to produce an answer. Many questions can be answered once a supporting object, action, or frame is localized, whereas others require additional reasoning, such as comparing events across time or composing multiple visual cues. This creates an *asymmetry*: evidence grounding is broadly necessary, but additional reasoning should be invoked only when grounded evidence is insufficient.

A common approach is to elicit textual chain-of-thought (CoT) rationales before the answer (Wei et al., 2022; Feng et al., 2025; Wang et al., 2025a). These rationales are especially useful when a question requires reasoning or multi-step inference, because they provide additional test-time computation. (Snell et al., 2024). However, this response format forces evidence localization and reasoning to be expressed in the same natural-language trace. For questions whose answer follows directly from a localized visual cue, the model still needs to ground that cue, but a full verbalized rationale adds decoding cost without providing additional visual evidence. This motivates a more efficient response format that stores query-relevant evidence in latent states and invokes additional reasoning only when the evidence must be composed.

Latent reasoning aims to provide such intermediate computation without verbalizing every step: instead of generating discrete rationale tokens, the model performs additional steps in continuous hidden states (Hao et al., 2025; Shen et al., 2025). Recent multimodal work further suggests that visual information can also be represented in latent space for perception-intensive reasoning (Li et al., 2025; Dong et al., 2026; Zhang et al., 2025b). However, video QA(VQA) poses a different challenge from image-centric latent reasoning, and a key question is how to organize these latent steps. A model must first localize query-relevant evidence across

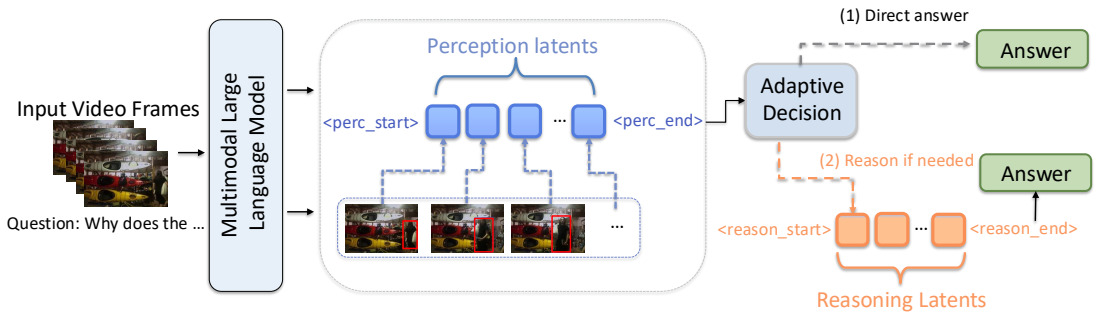

Figure 1: Overview of DyLaR, an evidence-first latent response format for video question answering. The model first grounds the question with perception latents that encode query-relevant visual evidence. It then adaptively chooses between a direct answer path and a reasoning path that inserts additional reasoning latents before the final answer. By allocating reasoning latents only when needed, DyLaR avoids unnecessary textual chain-of-thought generation while preserving answer-relevant visual evidence.

frames and then decide whether the localized evidence directly determines the answer or requires temporal, relational, or causal reasoning. A fixed, single-purpose latent segment used identically for all questions does not express this distinction. We therefore separate a common latent grounding stage from a conditional latent reasoning stage.

We propose Dynamic Latent Reasoning (**DyLaR**), which first grounds a question in latent visual evidence and then adaptively decides whether additional latent reasoning is needed. Given a video and a question, DyLaR first generates a short segment of *perception latents* that carries query-relevant visual evidence. The next generated token then serves as an adaptive decision point: the model either answers directly from the perception latents or extends the sequence with *reasoning latents* that provide latent reasoning before answering. Rather than assigning the same reasoning budget to every question, DyLaR uses perception latents as the evidence state and conditionally allocates reasoning latents only when further evidence reasoning is needed.

Learning response format requires two phases to acquire different functions. We use two supervision signals from verified video annotations: localized evidence boxes to provide object-level targets for perception latents, and verified rationales to supervise reasoning latents through rationale-to-latent distillation. Since each additional latent step adds decoding computation, the model is trained to allocate reasoning latents only for examples verified as requiring reasoning. We further refine complete responses with reinforcement learning under verifiable format and answer rewards.

Experiments show that DyLaR improves the accuracy–efficiency trade-off for video QA. On the Qwen3-VL-4B backbone (Bai et al., 2025a), DyLaR improves average accuracy across seven benchmarks over Qwen3-VL-4B-Thinking from 55.1 to 56.7 while generating only 16.6 tokens per query, compared with 2,431 tokens from Qwen3-VL-4B-Thinking and 6,950 tokens from explicit CoT prompting. On the Qwen2.5-VL-7B backbone (Bai et al., 2025b), DyLaR outperforms Video-R1 (Feng et al., 2025) and VideoRFT (Wang et al., 2025a) on average and remains close to Open-o3-Video (Meng et al., 2025), while using only 17.0 tokens per query. Ablations show that removing perception grounding, reasoning latents, or rationale-to-latent supervision all reduces accuracy, confirming that both grounded evidence states and latent evidence composition are important.

Our contributions are summarized as follows:

- We formulate VQA as evidence-first adaptive reasoning, where visual grounding is always required but additional reasoning is invoked only when needed. The response first grounds the question with perception latents, and then chooses between direct answering and latent reasoning before producing the answer.
- We introduce DyLaR, a latent response framework that separates perception latents for grounded visual evidence from optional reasoning latents for evidence reasoning.
- We train these latent phases with localized evidence supervision, rationale-to-latent distillation, and verifiable reinforcement learning, and show strong accuracy–efficiency trade-offs across seven video benchmarks.

2 Related Work

Video reasoning with MLLMs. Video MLLMs have progressed from general video dialogue and instruction tuning, such as Video-ChatGPT, Video-LLaVA, LLaVA-Video, and VideoLLaMA3, to explicit video reasoning systems (Maaz et al., 2024; Lin et al., 2024; Zhang et al., 2024, 2025a). Recent reasoning-oriented methods use CoT supervision and reinforcement learning to improve video understanding (Feng et al., 2025; Wang et al., 2025a), while Open-o3-Video grounds reasoning with explicit timestamps and bounding boxes (Meng et al., 2025). Other work explores frame-aware or agentic reasoning (Wang et al., 2024; Zhi et al., 2025; Ge et al., 2025; He et al., 2025; Wang et al., 2025b; Zou et al., 2026). However, these methods often rely on long textual reasoning traces, explicit evidence-gathering procedures, or additional interaction steps.

Latent reasoning. Explicit CoT improves reasoning but introduces token overhead (Wei et al., 2022). Latent reasoning instead moves intermediate computation into continuous hidden states (Chen et al., 2025). Coconut feeds hidden states back as continuous thoughts, and CODI compresses explicit CoT into continuous states through self-distillation (Hao et al., 2025; Shen et al., 2025). SoftCoT, hybrid latent reasoning, and multimodal continuous thoughts further explore soft or continuous alternatives to discrete CoT generation (Xu et al., 2025; Yue et al., 2025; Pham and Ngo, 2025).

Latent visual reasoning. Recent work extends latent reasoning to visual representations and has shown that continuous visual tokens or latent visual thoughts can improve fine-grained visual reasoning (Bigverdi et al., 2024; Li et al., 2025; Wu et al., 2026; Qin et al., 2025; Yang et al., 2025; Liu et al., 2026). LVR reconstructs query-relevant visual tokens in latent space, reducing the reliance on textual CoT for perception-intensive visual QA (Li et al., 2025). ILVR further interleaves text generation with latent visual states supervised by teacher-selected visual features from helper images (Dong et al., 2026). Latent Sketchpad (Zhang et al., 2025b) also explores visual latent generation as an internal sketchpad, enabling models to interleave textual reasoning with generated visual latents that can be decoded into interpretable sketches.

However, a limitation remains when these paradigms are applied to video QA. Existing latent visual reasoning methods primarily operate in image-centric or helper-image settings that rely on a single image. Video QA poses a different form of adaptivity: the model must first identify question-relevant evidence across multiple frames, but only some questions require that evidence to be further composed through temporal, relational, or causal reasoning. Some video questions can be answered using localized cues in a few frames, whereas others require reasoning about events over time. Thus, applying latent visual reasoning to videos requires more than inserting visual latents into a text reasoning trace; it calls for a response space that separates visual grounding from the optional need for additional reasoning.

3 Method

We introduce **DyLaR**, an evidence-first latent response framework that unifies visual grounding and conditional reasoning within a single autoregressive response. A *latent response* contains ordinary output tokens for segment delimiters (special tokens) and the final answer, together with continuous latent positions that carry intermediate hidden-state computation. These latent positions are serialized by placeholder tokens but are decoded in the model's hidden space rather than through explicit vocabulary tokens.

Given a video and a question, the model first grounds the query in the visual input using a small number of *perception latents*. Conditioned on these latents, it then either produces the answer directly or, when the question demands further inference, inserts *reasoning latents* that integrate the grounded evidence before answering. This separates common need for visual grounding from the question-dependent need for additional reasoning.

We describe adaptive latent response generation in Section 3.1, and then detail how supervised learning and reinforcement learning optimize this response space in Section 3.2.

3.1 Adaptive Latent Response Generation

Given a video-question pair (V, q), the base VLM encodes the sampled frames and the question into a multimodal prefix. Conditioned on this prefix, DyLaR generates a structured latent response: delimiter and answer tokens are generated as ordinary text tokens, while perception and reasoning steps are represented by continuous latent positions.

Latent-segment decoding. Each latent segment occupies a contiguous span of positions in the output sequence, marked by a fixed number of <|latent_pad|> placeholder symbols. These placeholders reserve sequence positions but carry no semantic content of their own. Inside such a segment, decoding deviates from standard autoregressive generation: instead of sampling a vocabulary token at each step and feeding its token embedding as the next input, the model takes the previous step's decoder hidden state and feeds it directly as the next input embedding. Computation therefore proceeds entirely in the continuous hidden-state space, without projecting back through the vocabulary. Once the predetermined number of latent steps has elapsed, decoding reverts to ordinary token-by-token generation: the model samples a discrete text token, embeds it normally, and continues as usual. We refer to this procedure as *latent-segment decoding*.

Response structure. We define a perception segment, a reasoning segment, an answer field as

$$\begin{aligned}\mathcal{P}_{K_p} &= [\texttt{<perc_start>}, \mathbf{z}^p_{1:K_p}, \texttt{<perc_end>}],\\ \mathcal{R}_{K_r} &= [\texttt{<reason_start>}, \mathbf{z}^r_{1:K_r}, \texttt{<reason_end>}],\\ \mathcal{A}(a) &= [\texttt{<answer>}, a, \texttt{</answer>}].\end{aligned}$$

Here, $\mathbf{z}^p_{1:K_p}$ denotes K_p perception latents, $\mathbf{z}^r_{1:K_r}$ denotes K_r reasoning latents, and a denotes the final answer. The placeholders inside both latent segments are written as <|latent_pad|> in the serialized response; their roles are determined by the surrounding delimiters.

Adaptive routing. The response therefore takes one of two forms:

$$\begin{aligned}\mathcal{S}^{\text{DIRECT}} &= [\mathcal{P}_{K_p}, \mathcal{A}(a)],\\ \mathcal{S}^{\text{REASON}} &= [\mathcal{P}_{K_p}, \mathcal{R}_{K_r}, \mathcal{A}(a)].\end{aligned} \tag{1}$$

The position after <perc_end> serves as the adaptive decision point: conditioned on the question and the generated perception latents, the model either emits <answer> to answer from grounded visual evidence, or emits <reason_start> and allocates K_r reasoning latents to integrate that evidence before answering.

3.2 Learning Grounded Perception and Conditional Reasoning

We optimize DyLaR in two stages. Supervised learning shapes what each latent segment represents: perception latents are aligned with localized visual evidence, while reasoning latents are trained to capture reasoning information. Reinforcement learning then refines complete responses using rewards for response format and answer correctness.

Training annotations. The supervised losses below require two annotation types: localized evidence boxes for perception grounding and rationales for reasoning-latent supervision. Existing video QA datasets rarely provide both types of annotations, so we augment video-QA examples with visual evidence and, where appropriate, rationales. For each video-question pair, we uniformly sample frames and use Qwen3.5-397B-A17B (QwenTeam, 2026) as an annotation generator to produce candidate annotations. For direct examples, the annotation contains the answer and evidence boxes that localize the supporting visual content; for reasoning examples, it additionally contains a step-by-step rationale. We then use Kimi-K2.5 (KimiTeam, 2026) as an independent verifier to filter these candidates, checking evidence relevance, answer support from the video, and rationale faithfulness for reasoning examples. After verification, examples without a required rationale are labeled DIRECT, while examples with both verified visual evidence and a verified rationale are labeled REASON. The annotation and verification prompt templates are provided in Appendix A.

Supervised Grounding and Rationale Distillation The supervised objective separates the functions of the two latent types. Perception latents are aligned with localized visual evidence, while reasoning latents are distilled from explicit rationales for examples that require additional inference. Let $s \in \{\text{DIRECT}, \text{REASON}\}$ denote the verified data type. The SFT objective is

$$\begin{aligned}\mathcal{L}_{\text{SFT}} = \; & \mathcal{L}_{\text{text}} + \lambda_g \mathcal{L}_{\text{ground}}\\ & + \mathbb{I}[s = \text{REASON}]\left(\lambda_e \mathcal{L}_{\text{exp}} + \lambda_d \mathcal{L}_{\text{distill}}\right),\end{aligned} \tag{2}$$

where $\mathcal{L}_{\text{text}}$ is the standard language-modeling loss over non-latent text positions in the target response. It includes structural delimiters and final answer tokens, while <|latent_pad|> positions are ignored. The grounding term encourages perception-latent hidden states to align with object-level visual targets constructed from verified evidence boxes. For REASON examples, the explicit-rationale term applies language-modeling supervision to the rationale-and-answer view, and the distillation term aligns the latent-reasoning view with

the answer-prediction state of the explicit-rationale view. λ_g, λ_e, and λ_d are hyperparameters.

Perception grounding. For each verified evidence box, we construct an object-level visual target from the video patches it covers. Let $\mathcal{P}_i$ denote the set of patch indices inside the i-th evidence box, and let $\mathbf{x}_p^v \in \mathbb{R}^d$ be the projected visual embedding of patch p in the language-model hidden space. The target for the i-th perception latent is the box-averaged embedding

$$\mathbf{u}_i = \frac{1}{|\mathcal{P}_i|} \sum_{p \in \mathcal{P}_i} \mathbf{x}_p^v, \qquad \mathbf{u}_i \in \mathbb{R}^d. \tag{3}$$

We aggregate patches within each box because the evidence needed for video QA is usually an object, region, or action-related unit rather than an isolated patch. Evidence boxes are ordered chronologically across frames and then by their annotation order within each frame.

During supervised learning, the number of perception latents follows the number of verified boxes in the example. Let M_p be this number. We teacher-force the perception-latent positions with these visual targets: $\mathbf{u}_i$ is used as the input embedding at the corresponding perception-latent position, and the preceding hidden state is aligned to $\mathbf{u}_i$ through the grounding loss.

Let $\hat{\mathbf{z}}_i^p$ denote the decoder hidden state immediately preceding the i-th perception latent, the grounding loss is the average cosine distance

$$\mathcal{L}_{\text{ground}} = \frac{1}{M_p} \sum_{i=1}^{M_p} \left[1 - \cos\left(\hat{\mathbf{z}}_i^p, \mathbf{u}_i\right)\right]. \tag{4}$$

At inference time, evidence boxes are not available; perception latents are generated using the latent-segment decoding rule described in Section 3.1.

Rationale-to-latent self-distillation. Reasoning latents do not have direct visual targets. Building on CODI's self-distillation framework (Shen et al., 2025), we adapt explicit-to-implicit CoT distillation to the reasoning segment in our video latent response sequence. For each REASON example, we construct two shared-parameter views. The latent view follows the reasoning form in Eq. (1), where the rationale is replaced by K_r reasoning latents. The explicit-rationale view uses the verified rationale before predicting the same answer. The explicit-rationale view is trained with language-modeling loss on its rationale and answer tokens, denoted by $\mathcal{L}_{\text{exp}}$.

To transfer rationale information into continuous states, we align the two views at the answer-prediction position. Let $\mathbf{h}_\ell^{\text{lat}}$ and $\mathbf{h}_\ell^{\text{exp}}$ be the layer-ℓ hidden states at this position in the latent and explicit-rationale views. We define

$$\mathcal{L}_{\text{distill}} = \frac{1}{L} \sum_{\ell=1}^{L} \text{SmoothL1}\left(\mathbf{h}_\ell^{\text{lat}}, \text{sg}\left[\mathbf{h}_\ell^{\text{exp}}\right]\right), \tag{5}$$

where L is the number of decoder layers and $\text{sg}[\cdot]$ denotes stop-gradient. SmoothL1 is applied element-wise and averaged over hidden dimensions: for $\Delta = \mathbf{x} - \mathbf{y}$, $\text{SmoothL1}(\mathbf{x}, \mathbf{y}) = \frac{1}{d}\sum_{j=1}^{d} \phi(\Delta_j)$, where $\phi(u) = \frac{1}{2}u^2$ if $|u| < 1$ and $\phi(u) = |u| - \frac{1}{2}$ otherwise. The latent view is not trained to imitate rationale tokens; the rationale provides representation-level supervision for the answer-relevant reasoning state.

Reinforcement Learning with Latent Replay SFT provides latent supervision under fixed target sequences, while inference depends on the model's own routing and final answer decisions. We therefore refine complete rollouts with GRPO using verifiable rewards for response-format validity and answer correctness.

Since latent tokens are continuous hidden states rather than sampled vocabulary tokens, they do not have token-level probabilities. We adopt latent-state replay following $\text{GRPO}_{\text{latent}}$ (Li et al., 2025).

For each policy-scored text token $y_{i,t}$ in rollout o_i, the importance ratio is

$$r_{i,t}(\theta) = \frac{\pi_\theta\left(y_{i,t} \mid V, q, \tilde{\mathbf{H}}_i, y_{i,<t}\right)}{\pi_{\theta_{\text{old}}}\left(y_{i,t} \mid V, q, \tilde{\mathbf{H}}_i, y_{i,<t}\right)}. \tag{6}$$

The token after <perc_end> is also policy-scored: <answer> yields the direct form, while <reason_start> yields the reasoning form.

For each prompt, we sample a group of G rollouts $\{o_i\}_{i=1}^{G}$ from the old policy. The clipped GRPO objective is

$$J_{\text{RL}}(\theta) = \mathbb{E}\Bigg[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|\mathcal{T}_i|}\sum_{t\in\mathcal{T}_i}\min\Big(r_{i,t}(\theta)\hat{A}_i, \text{clip}\big(r_{i,t}(\theta), 1-\epsilon, 1+\epsilon\big)\hat{A}_i\Big) - \beta D_{\text{KL}}\left(\pi_\theta \| \pi_{\text{ref}}\right)\Bigg]. \tag{7}$$

where $\mathcal{T}_i$ is the set of policy-scored text tokens in rollout o_i, $\pi_{\theta_{\text{old}}}$ is the old policy, π_{ref} is the frozen SFT model, ϵ is the clipping coefficient, β is the

Table 1: Main results on seven video benchmarks with 16-frame uniform sampling. **DyLaR** replaces long textual CoT with compact latent computation and generates only about 17 tokens per query on average. Within each backbone group, the best score per column is in bold and the second-best is underlined. Gen Tok. denotes the average number of generated tokens, LongVB denotes the LongVideo Bench.

Model	VMME-S	VMME-M	VMME-L	LVBench	MVBench	MMVU	LongVB	**Avg**	**Gen Tok.**
Proprietary models									
GPT-4o (1fps)	80.0	70.3	65.3	48.9	–	67.4	66.7	66.4	–
Gemini-1.5-Pro (1 fps)	81.7	74.3	67.4	33.1	–	65.4	64.0	64.3	–
Qwen2.5-VL-7B backbone									
Qwen2.5-VL-7B-Instruct + CoT	55.8	41.7	36.6	24.1	54.9	57.6	35.9	43.8	358
Video-R1	67.0	<u>56.6</u>	**50.3**	35.4	**68.3**	63.0	53.4	56.4	419
VideoRFT	67.7	54.9	<u>50.0</u>	34.9	67.2	**65.6**	53.3	56.2	361
Open-o3-Video	**70.1**	**58.2**	49.6	<u>37.5</u>	66.1	<u>65.1</u>	**57.1**	**57.7**	<u>154</u>
DyLaR (Qwen2.5-VL-7B)	<u>69.3</u>	55.3	48.9	**39.3**	66.9	63.8	<u>54.5</u>	<u>56.9</u>	**17.0**
Qwen3-VL-4B backbone									
Qwen3-VL-4B-Instruct + CoT	63.4	46.3	38.0	29.4	53.3	62.7	41.8	47.9	6950
Qwen3-VL-4B-Thinking	<u>67.9</u>	<u>52.4</u>	**48.2**	<u>36.7</u>	<u>56.9</u>	**66.7**	**56.4**	<u>55.1</u>	<u>2431</u>
DyLaR (Qwen3-VL-4B)	**71.1**	**55.1**	<u>46.6</u>	**40.5**	**61.5**	<u>65.4</u>	**56.4**	**56.7**	**16.6**

KL coefficient, and $\hat{A}_i$ is the group-normalized advantage.

Reward. We use a format-gated reward over complete rollouts:

$$R(o) = \mathbb{I}_{\text{form}}(o)\left(\alpha_{\text{form}} + \alpha_{\text{acc}}\mathbb{I}_{\text{acc}}(o)\right), \quad (8)$$

where $\mathbb{I}_{\text{form}}(o)$ checks whether the rollout follows valid response form in Eq. (1), and $\mathbb{I}_{\text{acc}}(o)$ checks whether the extracted answer matches the ground truth. The coefficients α_{form} and α_{acc} control the weights of format validity and answer accuracy.

4 Experiments

4.1 Experiment Details

Models. Our main model is trained from Qwen3-VL-4B-Instruct (Bai et al., 2025a). To compare with existing Qwen2.5-VL-based video reasoning baselines, we also instantiate DyLaR on Qwen2.5-VL-7B-Instruct (Bai et al., 2025b). For both backbones, we freeze the visual encoder and multimodal projector, and update only the language model parameters.

Training data. We curate training data primarily from Video-R1-CoT (Feng et al., 2025), which aggregates LLaVA-Video-178k (Zhang et al., 2025c), NeXT-QA (Xiao et al., 2021), PerceptionTest (Patraucean et al., 2023), etc., and supplement it with Video-Holmes (Cheng et al., 2025) for additional reasoning-heavy video questions. Following the annotation generation and verification procedure described in Section 3.2 and Appendix A, we retain approximately 20K examples for SFT and 30K prompts for reinforcement learning.

Training and inference. During SFT, we train for one epoch with AdamW at a learning rate of 1×10^{-5}. The number of perception latent tokens follows the number of verified evidence boxes in each SFT example, while the number of reasoning latent tokens is fixed at $K_r = 6$. During RL, we use GRPO with learning rate 1×10^{-6}, group size $G = 8$, sampling temperature $\tau = 0.9$, and KL coefficient $\beta = 0.04$. For both training and inference, we uniformly sample 16 frames per video, with each frame capped at 307,200 pixels. At inference time, we use greedy decoding with a fixed perception latent tokens budget $K_p = 4$ and reasoning latent tokens budget $K_r = 6$. We train our main Qwen3-VL-4B model for 2,500 RL steps. Due to compute constraints, the Qwen2.5-VL-7B variant is trained for 1,500 RL steps and is used mainly for comparison with Qwen2.5-VL-based baselines.

Benchmarks. We evaluate on video-QA benchmarks covering both perception-centric and reasoning-centric questions: the three duration splits of Video-MME (Fu et al., 2025) (short / medium / long), LVBench (Wang et al., 2025c), LongVideoBench (Wu et al., 2024), MVBench (Li et al., 2024), and MMVU (Zhao et al., 2025).

Baselines. We compare with three groups of baselines. First, we include reported results from proprietary models, GPT-4o and Gemini-1.5-Pro, as reference points for frontier model performance; these models may use different inference settings. Second, for the Qwen2.5-VL-7B backbone (Bai et al., 2025b), we compare against Qwen2.5-VL-7B-Instruct with CoT prompting and recent video reasoning models built on the same backbone, includ-

Table 2: SFT objective ablation on LVBench using Qwen3-VL-4B before RL. Each column reports accuracy on a LVBench category. "Perception latent + text CoT" keeps the perception latent block but replaces reasoning latents with explicit textual rationales. "w/o reasoning latents" removes the reasoning block. "w/o reasoning supervision" keeps the perception and reasoning latent blocks but removes $\mathcal{L}_{\text{exp}}$ and $\mathcal{L}_{\text{distill}}$. "w/o perception grounding" removes $\mathcal{L}_{\text{ground}}$.

Variant	Entity	Event	Key Information	Reason	Temporal	Summarization	**Overall**
Text-reasoning SFT baselines							
Text-CoT SFT	39.00	39.26	**42.27**	37.81	34.55	31.03	38.67
Perception latent + text CoT	39.00	36.63	39.52	32.84	33.18	**43.10**	37.57
DyLaR latent SFT							
DyLaR SFT full	**41.06**	**39.72**	40.89	**38.81**	**38.64**	36.21	**40.15**
w/o reasoning latents	39.73	36.32	34.02	38.81	37.73	32.76	37.51
w/o reasoning supervision	39.29	38.64	39.86	38.81	35.00	37.93	38.99
w/o perception grounding	38.70	38.79	40.89	35.82	35.91	32.76	38.61

ing Video-R1 (Feng et al., 2025), VideoRFT (Wang et al., 2025a), and Open-o3-Video (Meng et al., 2025). Third, for the Qwen3-VL-4B backbone, we compare against Qwen3-VL-4B-Instruct (Bai et al., 2025a) with explicit CoT prompting and Qwen3-VL-4B-Thinking (Bai et al., 2025a). For models we evaluate locally, we use the same 16-frame uniform sampling protocol.

4.2 Main Results

Table 1 reports the main results on seven video benchmarks. The proprietary models provide a scale reference, while the Qwen2.5-VL-7B and Qwen3-VL-4B groups compare models built on the same or closely related backbone families.

On the Qwen3-VL-4B backbone, DyLaR achieves the best average performance among the compared Qwen3 models, improving over Qwen3-VL-4B-Thinking by 1.6 points on average. The gains are especially clear on Video-MME Short (+3.2), Video-MME Medium (+2.7), LVBench (+3.8), and MVBench (+4.6). Although Qwen3-VL-4B-Thinking remains stronger on Video-MME Long and MMVU, DyLaR achieves the best overall average while generating only 16.6 visible tokens per query, compared with 2,431 tokens from Qwen3-VL-4B-Thinking and 6,950 tokens from explicit CoT prompting.

On the Qwen2.5-VL-7B backbone, DyLaR remains competitive with recent video reasoning models despite fewer RL steps. It outperforms Video-R1 and VideoRFT on average, achieves the best LVBench score within this backbone group, and trails Open-o3-Video by 0.8 points in accuracy. Meanwhile, DyLaR uses only 17.0 visible tokens on average, compared with 154 for Open-o3-Video, 419 for Video-R1, and 361 for VideoRFT. Qualitative examples visulization can be found in Appendix B.

Overall, DyLaR provides a favorable accuracy–efficiency trade-off: it matches or surpasses strong open-source reasoning baselines in backbone-controlled comparisons while replacing long reasoning traces with compact latent computation.

4.3 Ablation Studies and Analysis

SFT objective ablation. Table 2 studies which supervised signals are needed to initialize the adaptive text-latent response. We first compare against text-rationale SFT baselines. Text-CoT SFT performs reasoning entirely in the vocabulary space, while Perception latent + text CoT adds perception latents but still verbalizes the rationale. Both are below the full DyLaR SFT on LVBench overall, suggesting that simply using explicit rationales at inference does not provide the same compact latent response behavior.

Within the DyLaR family, removing reasoning latents reduces the overall score from 40.15 to 37.51, showing that visual grounding alone is insufficient for long-form video questions that require composing evidence across frames. Keeping the latent response structure but removing rationale supervision also lowers performance to 38.99, indicating that reasoning latents require explicit rationale-to-latent supervision rather than merely extra latent computation. Removing perception grounding further decreases overall accuracy to 38.61, confirming the importance of grounding perception latents in localized visual evidence.

RL refinement. Table 3 compares the full SFT checkpoint with the final RL-refined model on the Qwen3-VL-4B backbone. RL improves the average accuracy from 55.15 to 56.66, with clear gains

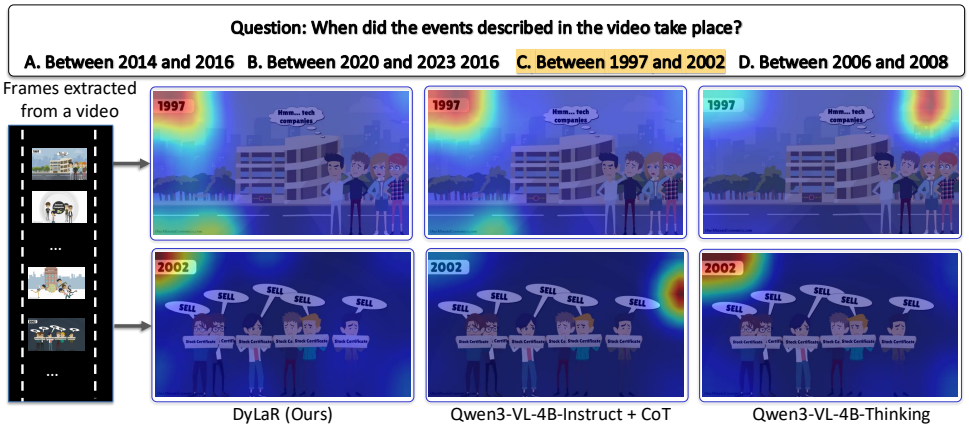

Figure 2: Case analysis and visualization. Heatmaps show Gaussian-smoothed attention aggregation over video patches for a temporal QA example. The example requires identifying the event period from two visual cues, "1997" and "2002". DyLaR focuses on the diagnostic temporal cues across frames, while CoT-based baselines show more diffuse or shifted attention.

Table 3: Effect of RL on Qwen3-VL-4B. Gen. Tok. denotes the average number of visible generated tokens.

Stage	LVBench	MVBench	MMVU	Mean	Gen. Tok.
SFT full	40.15	58.42	61.92	55.15	17.2
SFT + RL (DyLaR)	**40.48**	**61.50**	**65.44**	**56.66**	**17.1**
Δ	+0.33	+3.08	+3.52	+1.51	-0.1

on Video-MME Medium, MVBench, and MMVU, while keeping the visible generation length nearly unchanged. This suggests that supervised learning establishes the latent grounding and rationale-distillation behavior, and RL further improves the final text-latent response under verifiable rewards.

Adaptive routing analysis. Figure 3 analyzes whether DyLaR allocates reasoning only when needed. Using the official Video-MME category taxonomy, we measure the fraction of examples on which the model chooses the reasoning branch, i.e., generates reasoning latents before the answer. DyLaR triggers latent reasoning on only 32.9% of examples on average across seven perception categories, but on 82.4% of examples across four reasoning categories. This gap indicates that the model does not simply overuse the reasoning path; instead, it learns to reserve additional latent computation for categories that are more reasoning-intensive, consistent with the evidence-first adaptive design of DyLaR.

Case analysis and visualization. Figure 2 visualizes a representative temporal video QA example. The question asks when the events in the video took place, and the correct answer requires combining two visible temporal cues, "1997" and "2002",

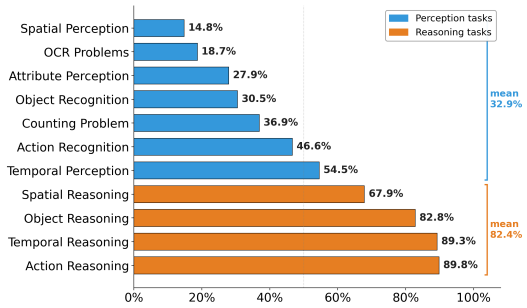

Figure 3: We report the fraction of examples on which DyLaR chooses the reasoning branch (generates reasoning latents before answering).

across different frames. Compared with Qwen3-VL-4B-Instruct CoT and Qwen3-VL-4B-Thinking, DyLaR produces more concentrated heatmaps on the diagnostic year cues and their supporting event regions. This suggests that the perception and reasoning latents help preserve localized visual evidence before the final answer decision, instead of relying solely on long textual CoT generation.

5 Conclusion

We presented DyLaR, an evidence-first latent response framework that grounds video questions with perception latents and conditionally allocates reasoning latents before answering. The model learns this response format from localized evidence supervision, rationale-to-latent distillation, and reinforcement learning with verifiable rewards. Across seven video benchmarks, DyLaR improves the accuracy–efficiency trade-off, and ablations confirm the value of both grounded perception states and latent reasoning.

Limitations

Although DyLaR reduces visible text generation by moving intermediate computation into latent states, these states are not directly human-readable, which limits the interpretability of individual reasoning steps. In addition, our experiments focus on Qwen-VL backbones and 16-frame uniform sampling; future work could study whether the same grounding–reasoning separation transfers to a broader set of architectures and adaptive frame-selection settings.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025a. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025b. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Mahtab Bigverdi, Zelun Luo, Cheng-Yu Hsieh, Ethan Shen, Dongping Chen, Linda G. Shapiro, and Ranjay Krishna. 2024. Perception tokens enhance visual reasoning in multimodal language models. *Preprint*, arXiv:2412.03548.

Xinghao Chen, Anhao Zhao, Heming Xia, Xuan Lu, Hanlin Wang, Yanjun Chen, Wei Zhang, Jian Wang, Wenjie Li, and Xiaoyu Shen. 2025. Reasoning beyond language: A comprehensive survey on latent chain-of-thought reasoning. *Preprint*, arXiv:2505.16782.

Junhao Cheng, Yuying Ge, Teng Wang, Yixiao Ge, Jing Liao, and Ying Shan. 2025. Video-holmes: Can mllm think like holmes for complex video reasoning? *arXiv preprint arXiv:2505.21374*.

Shuai Dong, Siyuan Wang, Xingyu Liu, Chenglin Li, Haowen Hou, and Zhongyu Wei. 2026. Interleaved latent visual reasoning with selective perceptual modeling. *Preprint*, arXiv:2512.05665.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. 2025. Video-r1: Reinforcing video reasoning in mllms. In *Advances in Neural Information Processing Systems*, volume 38, pages 99114–99137. Curran Associates, Inc.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, and 1 others. 2025. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24108–24118.

Haonan Ge, Yiwei Wang, Kai-Wei Chang, Hang Wu, and Yujun Cai. 2025. Framemind: Frame-interleaved video reasoning via reinforcement learning. *arXiv preprint arXiv:2509.24008*.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. 2025. Training large language models to reason in a continuous latent space. *Preprint*, arXiv:2412.06769.

Zefeng He, Xiaoye Qu, Yafu Li, Siyuan Huang, Daizong Liu, and Yu Cheng. 2025. Framethinker: Learning to think with long videos via multi-turn frame spotlighting. *arXiv preprint arXiv:2509.24304*.

KimiTeam. 2026. Kimi k2.5: Visual agentic intelligence. *Preprint*, arXiv:2602.02276.

Bangzheng Li, Ximeng Sun, Jiang Liu, Ze Wang, Jialian Wu, Xiaodong Yu, Hao Chen, Emad Barsoum, Muhao Chen, and Zicheng Liu. 2025. Latent visual reasoning. *arXiv preprint arXiv:2509.24251*.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, and 1 others. 2024. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2024. 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.

Chengzhi Liu, Yuzhe Yang, Yue Fan, Qingyue Wei, Sheng Liu, and Xin Eric Wang. 2026. Reasoning within the mind: Dynamic multimodal interleaving in latent space. *Preprint*, arXiv:2512.12623.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. 2024. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12585–12602.

Jiahao Meng, Xiangtai Li, Haochen Wang, Yue Tan, Tao Zhang, Lingdong Kong, Yunhai Tong, Anran Wang, Zhiyang Teng, Yujing Wang, and 1 others. 2025. Open-o3 video: Grounded video reasoning with explicit spatio-temporal evidence. *arXiv preprint arXiv:2510.20579*.

Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Mateusz Malinowski, Yi Yang, Carl Doersch, and 1 others. 2023. Perception test: A diagnostic benchmark for multimodal video models. *Advances in Neural Information Processing Systems*, 36:42748–42761.

Tan-Hanh Pham and Chris Ngo. 2025. Multimodal chain of continuous thought for latent-space reasoning in vision-language models. *Preprint*, arXiv:2508.12587.

Yiming Qin, Bomin Wei, Jiaxin Ge, Konstantinos Kallidromitis, Stephanie Fu, Trevor Darrell, and XuDong Wang. 2025. Chain-of-visual-thought: Teaching vlms to see and think better with continuous visual tokens. *Preprint*, arXiv:2511.19418.

QwenTeam. 2026. Qwen3.5: Accelerating productivity with native multimodal agents.

Zhenyi Shen, Hanqi Yan, Linhai Zhang, Zhanghao Hu, Yali Du, and Yulan He. 2025. Codi: Compressing chain-of-thought into continuous space via self-distillation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 677–693.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *Preprint*, arXiv:2408.03314.

Qi Wang, Yanrui Yu, Ye Yuan, Rui Mao, and Tianfei Zhou. 2025a. Videorft: Incentivizing video reasoning capability in mllms via reinforced fine-tuning. *Preprint*, arXiv:2505.12434.

Shijian Wang, Jiarui Jin, Xingjian Wang, Linxin Song, Runhao Fu, Hecheng Wang, Zongyuan Ge, Yuan Lu, and Xuelian Cheng. 2025b. Video-thinker: Sparking" thinking with videos" via reinforcement learning. *arXiv preprint arXiv:2510.23473*.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, and 1 others. 2025c. Lvbench: An extreme long video understanding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22958–22967.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. 2024. Videoagent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision*, pages 58–76. Springer.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

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

Linquan Wu, Tianxiang Jiang, Yifei Dong, Haoyu Yang, Fengji Zhang, Shichaang Meng, Ai Xuan, Linqi Song, and Jacky Keung. 2026. Lavit: Aligning latent visual thoughts for multi-modal reasoning. *Preprint*, arXiv:2601.10129.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. 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.

Yige Xu, Xu Guo, Zhiwei Zeng, and Chunyan Miao. 2025. Softcot: Soft chain-of-thought for efficient reasoning with llms. *Preprint*, arXiv:2502.12134.

Zeyuan Yang, Xueyang Yu, Delin Chen, Maohao Shen, and Chuang Gan. 2025. Machine mental imagery: Empower multimodal reasoning with latent visual tokens. *Preprint*, arXiv:2506.17218.

Zhenrui Yue, Bowen Jin, Huimin Zeng, Honglei Zhuang, Zhen Qin, Jinsung Yoon, Lanyu Shang, Jiawei Han, and Dong Wang. 2025. Hybrid latent reasoning via reinforcement learning. *Preprint*, arXiv:2505.18454.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, and 1 others. 2025a. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*.

Huanyu Zhang, Wenshan Wu, Chengzu Li, Ning Shang, Yan Xia, Yangyu Huang, Yifan Zhang, Li Dong, Zhang Zhang, Liang Wang, Tieniu Tan, and Furu Wei. 2025b. Latent sketchpad: Sketching visual thoughts to elicit multimodal reasoning in mllms. *Preprint*, arXiv:2510.24514.

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

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. 2025c. Llava-video: Video instruction tuning with synthetic data. *Preprint*, arXiv:2410.02713.

Yilun Zhao, Haowei Zhang, Lujing Xie, Tongyan Hu, Guo Gan, Yitao Long, Zhiyuan Hu, Weiyuan Chen, Chuhan Li, Zhijian Xu, and 1 others. 2025. Mmvu: Measuring expert-level multi-discipline video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8475–8489.

Zhuo Zhi, Qiangqiang Wu, Wenbo Li, Yinchuan Li, Kun Shao, Kaiwen Zhou, and 1 others. 2025. Videoagent2: Enhancing the llm-based agent system for

long-form video understanding by uncertainty-aware cot. *arXiv preprint arXiv:2504.04471*.

Junbo Zou, Ziheng Huang, Shengjie Zhang, Liwen Zhang, and Weining Shen. 2026. Videobrain: Learning adaptive frame sampling for long video understanding. *arXiv preprint arXiv:2602.04094*.

A Data Annotation and Verification Prompts

This appendix provides the prompt templates used to construct and verify the training annotations for DyLaR. The annotation pipeline is designed to provide two types of supervision: localized visual evidence for perception latents and, when additional inference is required, concise rationales for rationale-based latent supervision.

A.1 Annotation Overview

For each video-question pair, we uniformly sample 16 frames and use Qwen3.5-397B-A17B as the annotation generator. The generator produces a structured annotation containing a final answer, a set of question-relevant evidence boxes, and an optional rationale. We distinguish two answerable modes. In DIRECT examples, the localized evidence directly supports the answer. In REASON examples, the relevant evidence is visible but must be composed through temporal comparison, causal reasoning, relational reasoning, or option-level reasoning.

We then use Kimi-K2.5 as an independent verifier. The verifier checks whether the generated boxes contain useful evidence, whether the boxes support the gold answer, and, for reasoning examples, whether the generated rationale is faithful to the visible evidence. We discard malformed outputs, invalid boxes, unsupported samples, and annotations whose final answers disagree with the gold answer. Only verified examples are used for supervised fine-tuning.

A.2 Annotation Generation Prompt

The annotation generator is instructed to localize visual evidence before deciding whether the example should be treated as direct perception or reasoning. The simplified prompt template is shown below.

Annotation Generation Prompt

```
You are an evidence-first video annotation
assistant. Return strict JSON only.

Input:
- A video question.
- Candidate answer options, if available.
- 16 uniformly sampled video frames with
frame indices and timestamps.

Task:
1. Annotate all question-relevant visual
evidence in the visible frames.
  Evidence may include subjects, objects,
actions, states, transitions, temporal
bookends, or visual cues needed to answer
the question.
2. Decide whether the current visual
evidence is sufficient.
3. If the evidence directly supports the
answer, use mode = "perception_only".
  Provide evidence boxes, the final
answer, and an empty rationale.
4. If the evidence is visible but requires
comparison, temporal integration,
causality, relational reasoning, or option
elimination, use mode = "need_reasoning".
  Provide evidence boxes, the final
answer, and one concise rationale.
5. If the question is not answerable from
the visible evidence, use mode =
"unsupported".

Rules for evidence boxes:
- Each evidence item should localize a
concrete visual region.
- Use normalized box coordinates.
- Avoid unrelated background regions or
decorative objects.
- Multiple boxes and multiple frames are
allowed when they support the answer.

Rules for rationale:
- Use a rationale only for mode =
"need_reasoning".
- Write one concise sentence describing the
reasoning over grounded objects or events.
- Do not mention frame IDs, box IDs,
coordinates, or annotation metadata.
- Do not reveal the answer by saying "the
answer is A/B/C/D" inside the rationale.

Output strict JSON:
{
  "mode": "perception_only" |
"need_reasoning" | "unsupported",
  "evidence": [
    {
      "frame_id": integer,
      "timestamp": float,
      "entity": "short visual phrase",
      "bbox": [x1, y1, x2, y2],
      "role": "evidence role"
    }
  ],
  "rationale": "empty for
perception_only",
  "answer": "answer label or text"
}
```

A.3 Three-Stage Verification Prompts

We verify generated annotations with three lightweight checks. The first two phases operate at the evidence-box level, while the third phase verifies rationale faithfulness for REASON examples.

Phase A: Gold-blind crop usefulness. For each generated box, we crop the corresponding region with padding. The verifier sees the question and the cropped regions, but not the gold answer. This phase filters boxes that are empty, blurred, or irrelevant to the question.

Phase A Verification Prompt

```
Question: {question}

You are shown cropped image regions, each
labeled with a box ID.

For each crop, judge whether it contains
useful visual evidence for answering the
question.

Mark false only if:
1. the crop is empty, blurred, or contains
no question-relevant content;
2. the crop is unrelated to the question.

Otherwise mark true.

Output strict JSON:
{
  "judgments": [
    {
      "bbox_id": integer,
      "evidence_useful": true | false,
      "confidence": float between 0 and 1,
      "reason": "at most 20 words"
    }
  ]
}
```

Phase B: Answer support verification. In Phase B, the verifier sees the question, the gold answer, and the original frames with the Phase-A surviving boxes overlaid. This phase checks whether each box supports the correct answer in its frame context.

Phase B Verification Prompt

```
Question: {question}
Correct answer: {gold_answer}

You are shown video frames with numbered
bounding boxes.

For each numbered box, judge whether the
content inside the box directly supports
the correct answer.

Use the rest of the frame as spatial or
contextual reference if the question
involves position, comparison, or
relationships.

If a temporal question requires multiple
frames, give benefit of the doubt to a box
that contributes necessary partial
evidence.

Output strict JSON:
{
  "judgments": [
    {
      "bbox_id": integer,
      "supports_answer": true | false,
      "confidence": float between 0 and 1,
      "reason": "at most 20 words"
    }
  ]
}
```

Phase C: Rationale faithfulness. Phase C is applied only to REASON examples. The verifier receives the question, gold answer, generated rationale, and key frames with evidence boxes overlaid. It checks whether the rationale is logical, visually grounded, and free from answer leakage or hallucinated visual content.

Phase C Verification Prompt

```
Question: {question}
Correct answer: {gold_answer}
Rationale: {rationale}

Below are key video frames with evidence
boxes around relevant entities.

Judge whether the rationale faithfully
reflects the visible evidence.

Check:
1. "logical_to_answer": the reasoning
chain validly leads to the correct answer.
2. "visually_grounded": the entities,
actions, and events mentioned in the
rationale are visible in the provided
frames.
3. "leaks_answer_letter": the rationale
directly states the answer label, such as
"the answer is A".
4. "hallucinates_objects": the rationale
describes objects, actions, or events not
visible in the frames.

Output strict JSON:
{
  "logical_to_answer": true | false,
  "visually_grounded": true | false,
  "leaks_answer_letter": true | false,
  "hallucinates_objects": true | false,
  "reason": "at most 40 words"
}
```

The verified annotations determine which supervised signals are applied during training. Verified evidence boxes provide object-level targets for perception latents, while verified rationales are used only in the explicit-rationale view for rationale-to-latent distillation. The latent placeholder positions are not supervised as ordinary text tokens.

B More Visualizations

In this section, Figure 4 and Figure 5 show two routes of our DyLaR model. The model distinguished the task successfully and therefore entered the correct latent route.

Question: What color are the protective gloves worn by the person in the video?
A. Yellow and green. B. Red and yellow. C. White. D. Blue.
Ground Truth: B. Red and yellow.

Model Output:
<perc_start><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><perc_end><answer>B</answer>

Figure 4: DyLaR distinguished this case does not require reasoning, as the color of the glove is obviously shown in the image.

Question: According to the information presented in the video, what natural disaster did France experience that was identified as its worst ever?
A. drought. B. Sandstorms. C. Floods. D. Snowstorms.
Ground Truth: A. drought.

Model Output: <perc_start><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><perc_end>
<reason_start><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><reason_end><answer>A</answer>

Figure 5: DyLaR routes this case through its reasoning segment: the frames show its downstream wildfires—rather than the disaster itself. The model infers the visible evidence into the underlying cause (drought).