CALIBRATING PERCEPTION AND TEMPORAL TRAJECTORIES FOR TEST-TIME LATENT ALIGNMENT IN MULTIMODAL LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

While existing MLLMs enhance logical reasoning, they still suffer from severe confidence miscalibration. Moreover, current explicit and implicit reasoning mechanisms are constrained by discrete tokens and autoregressive mapping, which easily induces "premature semantic collapse" and "tunnel vision," failing to meet the holistic and hierarchical demands of visual perception.

To address these challenges, we propose CHARM, a progressive three-stage architecture: First, Metacognitive Visual Learning (MVL) uses reinforcement learning to jointly optimize accuracy, calibration, and sensitivity, thereby aligning perception with confidence. Second, Chronological Trajectory Refinement (CTR) introduces temporal perturbation penalties to reinforce explicit temporal reasoning and keyframe grounding. Finally, Multi-Horizon Latent Alignment (MHLA) operates during test-time inference, leveraging pre-calibrated uncertainty to dynamically mix future multi-step semantic superposition distributions with hard labels, optimizing latent reasoning to decode the final explicit textual answer.

Built on Qwen2.5-VL-7B, our method achieves 80.5% (OE) and 87.3% (MC) on Math-Vista, as well as 72.8% (STEM) and 80.6% (HASS) on MMMU. When scaled to Qwen3-VL-8B-Instruct, it reaches 69.7 (R1@0.3) on the ActivityNet benchmark and sets new state-of-the-art records on QVHighlights and TVGBench with 90.3 and 65.1 (R1@0.3).

1 INTRODUCTION

Recent advances in Multimodal Large Language Models (MLLMs) main focus on optimizing visual data distributions Hinton et al. (2006), visual position encodings Huang et al. (2025), and visual instruction tuning Cai et al. (2025) to improve visual perception and accuracy; however, experiments demonstrate that model confidence remains stable even under severe perception degradation, exposing a critical confidence miscalibration issue; although pure-text Large Language Models (LLMs) mitigate this problem through token-level uncertainty estimation Ji et al. (2025), such approaches suffer from a granularity mismatch with visual perception that manifests holistically across the entire response; to address this, recent research calculates the mean negative log-probability across all output tokens Fu et al. (2025a) to represent overall confidence, striving to bridge the gap between what models see and what they claim to know.

For enhancing the reasoning capabilities of Multimodal Large Language Models (MLLMs), several representative works have made key explorations: DeepVideo-R1 Park et al. (2026) applies reinforcement learning fine-tuning to boost reasoning in video question answering; One-Thinker Feng et al. (2025) proposes an all-in-one reasoning architecture to handle diverse visual tasks; Video-R1 Feng et al. (2026) introduces the T-GRPO mechanism to encourage temporal awareness by contrasting performance on ordered versus shuffled videos; meanwhile, foundational studies such as VQALS Yang et al. (2025a) and ToolVQA Yin et al. (2025) have jointly witnessed and driven the rapid evolution of MLLM reasoning capabilities.

Moreover, Vision-Language Models (VLMs) have revolutionized visual understanding by integrating Large Language Models (LLMs) with visual encoders OpenAI (2023), yet explicit Chain-of-Thought (CoT) Wei et al. (2022) remains constrained by an information bandwidth bottleneck

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

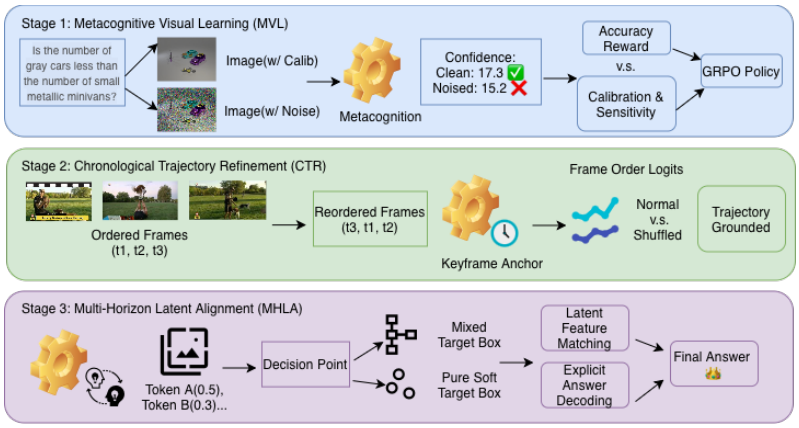


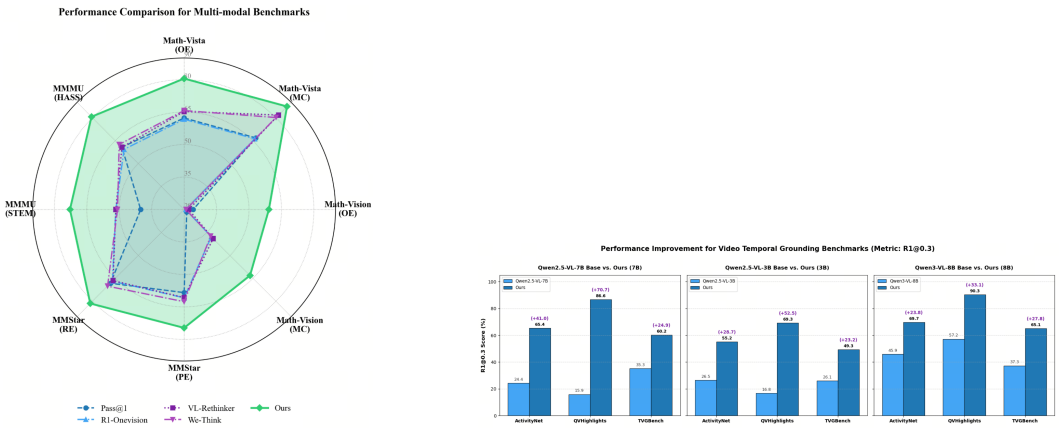
Figure 1: Overall Structure

caused by discrete tokenization Li et al. (2025). Although emerging latent space reasoning attempts to reason within high-dimensional hidden states, it retains autoregressive objectives that force strict point-wise mapping, contradicting the hierarchical nature of visual perception, which evolves from global semantics to local feature extraction Navon (1977). Forcing the latent state to prematurely ”collapse” into specific object tokens before grasping the holistic context induces premature semantic collapse, creating a ”tunnel vision” effect that hinders the capture of complex relationships.

Existing MLLMs face core challenges: confidence mis-calibration and granularity mismatches caused by visual perception degradation, developmental bottlenecks and adaptability constraints in complex visual and temporal reasoning, and semantic collapse with ”tunnel vision” resulting from discrete CoT bandwidth limitations alongside auto-regressive hidden-state mapping in latent space reasoning. Therefore, this study proposes CHARM (Calibrated Horizon Alignment for Reasoning in MLLMs) (Figure 1), a progressive three-stage architecture: Stage 1 (Metacognitive Visual Learning) utilizes GRPO Shao et al. (2024) with a joint reward function to optimize perception and confidence calibration, endowing the model with uncertainty estimation and perturbation-awareness capabilities; Stage 2 (Chronological Trajectory Refinement) employs reinforcement learning and a temporal frame-shuffling mechanism to enhance temporal sensitivity, guiding the generation of explicit reasoning paths anchored to keyframes; and Stage 3 (Multi-Horizon Latent Alignment) constructs look-ahead semantic windows during inference and adaptively aligns latent states via entropy-gated dynamic targets, ultimately integrating cross-entropy loss to achieve precise textual decoding.

Based on the radar chart across multimodal benchmarks (including Math-Vista Lu et al. (2023), Math-Vision Wang et al. (2024a), MMStar Chen et al. (2024), and MMMU Yue et al. (2024)) (Figure 2(a)), our method (Ours) occupies the largest area, consistently outperforming both the Pass@1 baseline and representative comparative methods such as R1-Onevision Yang et al. (2025c), VL-Rethinker Wang et al. (2025a), and We-Think Yang et al. (2025b) across all metrics. Specifically, Ours achieves new state-of-the-art performance on Math-Vista Lu et al. (2023) (OE: 80.5%, MC: 87.3%) and Math-Vision Wang et al. (2024a) (OE: 59.2%, MC: 63.1%), demonstrating a substantial margin of improvement. These results fully validate the powerful generalization ability and superior overall performance of our proposed method.

Besides, extensive evaluations across video temporal grounding benchmarks (including ActivityNet Krishna et al. (2017), QVHighlights Lei et al. (2021), and TVGBench Wang et al. (2026)) demonstrate the remarkable superiority of our method across various backbone scales (Figure 2(b)). On the QVHighlights dataset Lei et al. (2021), our approach achieves extraordinary performance gains: it not only elevates the Qwen2.5-VL-7B Bai et al. (2025b) baseline to 86.6% (an absolute increase of +70.7%), but also reaches a peak score of 90.3% (+33.1%) when integrated with Qwen3-VL-8B Bai et al. (2025a). These empirical results substantiate the robust generalization ability and strong scalability of our proposed framework across different model capacities.



(a) (b)

Figure 2: Overall Performance Analysis. (a) Radar Chart (base model: Qwen2.5-VL-7B Bai et al. (2025b)). Notes: OE (Open-Ended), MC (Multi-Choice), PE (Perception), RE (Reasoning), STEM (Science, Technology, Engineering, and Mathematics), HASS (Humanities, Arts, and Social Sciences). (b) Performance Improvement for Video Temporal Grounding Benchmarks (Metric: R1@0.3).

2 Related Work

Multi-modal Large Language Models (MLLMs) suffer from severe systemic miscalibration when evaluating their own outputs, and unlike mature research on LLM calibration, work on MLLM calibration is still nascent, facing the core challenge of addressing the unique impact of the visual component Geng et al. (2023). This issue first gains traction in high-risk domains such as clinical diagnostics and autonomous driving Du et al. (2025), and subsequent research further explores training-free methods applied to in-context learning for medical image classification Shen et al. (2025). As researchers come to identify hallucinations as an extreme form of calibration failure, advancements such as visual contrastive decoding Leng et al. (2023) and DPO-based alignment Compagnoni et al. (2025) emerge; however, these efforts largely remain incremental extensions of traditional LLM calibration methods, failing to adequately address how core concepts like visual perception fundamentally impact calibration outcomes.

In addition, TimeChat Ren et al. (2024) and UniVTG Lin et al. (2023) represent pioneering works that bring MLLMs into the Video Temporal Grounding (VTG) domain, successfully reframing traditional grounding tasks into text generation tasks that directly predict timestamps. To address the challenge of coarse temporal modeling, UniTime Li et al. (2026) introduces specialized time tokens to capture temporal information with higher precision. Subsequently, VideoChat-R1.5 Yan et al. (2026) and Time-R1 Wang et al. (2026) mark a major breakthrough by incorporating reinforcement learning (RL) into this field, leveraging a reasoning-guided post-training framework that encourages models to generate explicit thinking processes prior to prediction and achieve state-of-the-art performance.

Moreover, Large language models (LLMs) explore implicit thinking mechanisms entirely within latent states through Quiet-STaR Zelikman et al. (2024) to overcome the information loss caused by discrete tokenization in explicit Chain-of-Thought; subsequently, Coconut Hao et al. (2024) expands this paradigm, enabling models to perform continuous reasoning directly within the latent space. In the realm of vision-language models (VLMs), CoCoVa Ma et al. (2025) anchors latent thinking in visual evidence by utilizing latent attention to refine feature representations; meanwhile, Latent Sketchpad Zhang et al. (2025) constructs a ”visual scratchpad” that balances efficient latent-space reasoning with model interpretability through reconstructable latent sketches; building on this, advanced research represented by Monet Wang et al. (2025b) moves beyond simple supervised anchoring to directly optimize latent reasoning trajectories, further enhancing multimodal latent thinking capabilities.

3 METHODOLOGY

CHARM integrates three progressive stages: Stage 1 establishes perception and confidence alignment through Metacognitive Visual Learning (MVL); Stage 2 refines temporal reasoning paths using Chronological Trajectory Refinement (CTR); and Stage 3 executes test-time inference via Multi-Horizon Latent Alignment (MHLA).

3.1 STAGE 1: METACOGNITIVE VISUAL LEARNING (MVL)

In the first stage, the base model undergoes perception and confidence alignment training via Metacognitive Visual Learning (MVL), endowing the model with a metacognitive capacity regarding input perturbations and prediction certainty. Using Group Relative Policy Optimization (GRPO), we optimize the joint reward function:

$$\mathcal{R}_{\text{MVL}} = \mathcal{R}_{\text{acc}} + \lambda_{\text{calib}} \cdot \mathcal{R}_{\text{calib}} + \lambda_{\text{sens}} \cdot \mathcal{R}_{\text{sensitivity}} \tag{1}$$

where $\mathcal{R}_{\text{MVL}}$ represents the total reward for the MVL stage, $\mathcal{R}_{\text{acc}}$ denotes the task accuracy reward, $\mathcal{R}_{\text{calib}}$ is the confidence calibration reward that penalizes misjudgments stemming from over-confidence or under-confidence, $\mathcal{R}_{\text{sensitivity}}$ is the perception sensitivity reward enforcing a sensitive drop in output confidence under input perturbations, and λ_{calib} and λ_{sens} are scaling hyperparameters balancing the calibration and sensitivity reward components, respectively.

3.2 STAGE 2: CHRONOLOGICAL TRAJECTORY REFINEMENT (CTR)

Building upon the calibrated base model, the second stage employs Chronological Trajectory Refinement (CTR) to enhance temporal sensitivity and explicit reasoning path generation. Reinforcement learning is applied to guide and penalize sensitivity to temporal perturbations:

$$\mathcal{R}_{\text{CTR}} = \mathcal{R}_{\text{IoU/Acc}} + \gamma \cdot \Delta\text{Logit}_{\text{temporal}} \tag{2}$$

where $\mathcal{R}_{\text{CTR}}$ is the total reward for the CTR stage, $\mathcal{R}_{\text{IoU/Acc}}$ denotes the reward evaluating temporal grounding accuracy or intersection-over-union performance, $\Delta\text{Logit}_{\text{temporal}}$ measures the change in prediction logits under temporal frame shuffling to compel the model to generate reasoning trajectories anchored to spatial-temporal keyframes, and γ is a weighting hyperparameter controlling the influence of the temporal perturbation reward.

3.3 STAGE 3: MULTI-HORIZON LATENT ALIGNMENT (MHLA)

Following the completion of the first two training stages, the model enters the test-time decoding phase. We introduce Multi-Horizon Latent Alignment (MHLA) to dynamically superimpose and adaptively align future semantic horizons within the latent space.

At the t-th latent reasoning step, the hybrid target distribution Q_t is constructed over ungenerated future token semantics within a dynamic window W_t:

$$Q_t = \sum_{k \in W_t} w_k \cdot P_\theta(\cdot \mid I, Q, c_{<t}, k) \tag{3}$$

where Q_t denotes the constructed superposition distribution at the t-th latent step, W_t represents the dynamic look-ahead semantic window containing future tokens at step t, k is a candidate token within window W_t, w_k is the weighting coefficient assigned to token k, P_θ denotes the model parameterization, I is the input image or video context, Q is the user query, and $c_{<t}$ represents the sequence of latent contexts generated prior to step t.

Leveraging the reliable uncertainty derived from the first-stage MVL training (quantified via the model entropy $H(Q_t)$), the target distribution adaptively switches between or mixes the soft superposition target Q_t and the deterministic hard label y_{hard}:

$$P_t^{\text{target}} = \begin{cases} \alpha \cdot y_{\text{hard}} + (1-\alpha) \cdot Q_t, & \text{if } H(Q_t) > \eta \\ Q_t, & \text{otherwise} \end{cases} \tag{4}$$

where P_t^{target} is the adaptively constructed target probability distribution at step t, y_{hard} represents the deterministic ground-truth hard label distribution, Q_t is the soft superposition target distribution,

$H(Q_t)$ denotes the Shannon entropy of distribution Q_t serving as an uncertainty metric, η is a predefined entropy threshold for uncertainty gating, and α is a blending hyperparameter determining the proportion of the hard label used when uncertainty exceeds the threshold.

During the latent reasoning phase, the latent state h_t is aligned with the superimposed future semantic distribution:

$$\mathcal{L}_{\text{MHLA}} = -\frac{1}{T}\sum_{t=1}^{T}\sum_{k\in W_t} P_t^{\text{target}}(k)\log P_\theta(k \mid I, Q, c_{<t}) \tag{5}$$

where $\mathcal{L}_{\text{MHLA}}$ denotes the Multi-Horizon Latent Alignment loss, T is the total number of latent reasoning steps, W_t is the dynamic semantic window at step t, k represents an individual token within W_t, $P_t^{\text{target}}(k)$ is the target probability assigned to token k at step t, and $P_\theta(k \mid I, Q, c_{<t})$ is the model's predicted probability of token k given input I, query Q, and preceding latent context $c_{<t}$.

Upon completing the latent reasoning steps, the standard cross-entropy loss is employed to generate the final explicit textual answer:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{M}\sum_{j=1}^{M}\log P_\theta(a_j \mid I, Q, \mathcal{C}, a_{<j}) \tag{6}$$

where $\mathcal{L}_{\text{CE}}$ represents the explicit answer cross-entropy loss, M is the length of the final textual answer sequence, a_j is the j-th ground-truth answer token, $a_{<j}$ denotes the sequence of answer tokens generated prior to index j, $\mathcal{C}$ is the complete latent reasoning context accumulated during the latent steps, and $P_\theta(a_j \mid I, Q, \mathcal{C}, a_{<j})$ is the predicted probability of token a_j.

The total optimization objective for the inference and decoding phase is formulated as the combination of the MHLA loss and the cross-entropy loss:

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{MHLA}} + \mathcal{L}_{\text{CE}} \tag{7}$$

where $\mathcal{L}_{\text{Total}}$ represents the combined total optimization loss, $\mathcal{L}_{\text{MHLA}}$ is the Multi-Horizon Latent Alignment loss governing latent space reasoning, and $\mathcal{L}_{\text{CE}}$ is the cross-entropy loss governing explicit textual answer decoding.

4 Experiment

4.1 Experiment Settings

The complete experimental setup of the CHARM framework incorporates a progressive three-stage optimization and decoding mechanism. At the model and hardware level, Qwen2.5-VL-7B Bai et al. (2025b) serves as the main base model, with the vision encoder and modality merger frozen to fine-tune only the decoder parameters; an 8*A100 parallel computing setup is utilized alongside DeepSpeed ZeRO-3 and FlashAttention-2 to optimize training efficiency. In Stage 1 (Metacognitive Visual Learning, MVL), the GRPO algorithm Shao et al. (2024) is applied with an $8 \times G$ sampling strategy, injecting input perturbations to jointly optimize task accuracy, confidence calibration, and perception sensitivity rewards (with scaling weights $\lambda_{\text{calib}} = 0.5$ and $\lambda_{\text{sens}} = 0.3$). Built upon the calibrated model, Stage 2 (Chronological Trajectory Refinement, CTR) introduces a temporal frame-shuffling perturbation with $\Delta t = 3$s, enforcing temporal IoU and logit-change rewards ($\gamma = 0.3$) under a 2 FPS frame rate and a 3,584 video token limit. Stage 3 (Multi-Horizon Latent Alignment, MHLA) is applied during the test-time inference and decoding phase, leveraging the metacognitive entropy threshold $\eta = 0.6$ derived from Stage 1 to adaptively switch/blend ($\alpha = 0.8$) the superimposed soft target within the dynamic window W_t and the hard label, ultimately minimizing $\mathcal{L}_{\text{MHLA}} + \mathcal{L}_{\text{CE}}$ for end-to-end reasoning optimization.

For training datasets, Stage 1 (MVL) utilizes general VQA (RealWorldQA xAI Organization (2024), MMBench Liu et al. (2024), MMMU-Pro Yue et al. (2025)) and mathematical reasoning datasets (MathVerse Zhang et al. (2024), DynaMath Zou et al. (2025), LogicVista Xiao et al. (2024)) to perform GRPO-based Shao et al. (2024) perception and confidence alignment, enhancing model calibration and sensitivity under perturbations. Stage 2 (CTR) focuses on video temporal grounding

Table 1: Model performance comparison across multiple multi-modal benchmarks.

	Math-Vista Lu et al. (2023)		Math-Vision Wang et al. (2024a)		MMStar Chen et al. (2024)		MMMU Yue et al. (2024)	
Model/Dataset	**OE**	**MC**	**OE**	**MC**	**PE**	**RE**	**STEM**	**HASS**
Training Free Baselines (Qwen2.5-VL-7B Bai et al. (2025b))								
Pass@1	62.3	66.8	24.2	21.7	58.4	68.2	40.1	60.7
Majority Voting	68.4	73.7	26.2	33.2	63.6	77.0	53.2	62.5
Deepconf Fu et al. (2025b)	69.3	74.7	26.4	32.3	58.2	71.0	51.5	62.5
Training Based Framework								
R1-Onevision Yang et al. (2025c)	61.7	66.2	20.8	37.8	60.8	66.9	51.6	59.3
VL-Rethinker Wang et al. (2025a)	65.3	81.7	22.1	39.0	60.7	66.5	51.6	60.7
We-Think Yang et al. (2025b)	65.7	80.1	20.9	37.4	62.6	70.1	50.8	62.5
Ours	**80.5**	**87.3**	**59.2**	**63.1**	**74.7**	**81.4**	**72.8**	**80.6**
Training Free Baselines (Qwen3-VL-2B-Think Bai et al. (2025a))								
Pass@1	65.0	82.5	47.7	50.0	62.8	56.0	58.8	67.0
Majority Voting	72.3	89.9	60.9	62.5	67.1	66.5	72.9	81.0
DeepConf Fu et al. (2025b)	72.8	87.9	52.2	62.5	67.4	68.1	73.8	82.3
Ours	**82.9**	**93.7**	**65.4**	**71.2**	**75.5**	**81.3**	**82.5**	**88.1**

Table 2: Performance evaluation on video temporal grounding benchmarks.

		ActivityNet Krishna et al. (2017)			QVHighlights Lei et al. (2021)			TVGBench Wang et al. (2026)		
Method	**Size**	**R1@0.3**	**R1@0.5**	**R1@0.7**	**R1@0.3**	**R1@0.5**	**R1@0.7**	**R1@0.3**	**R1@0.5**	**R1@0.7**
Qwen2.5-VL-7B-Instruct Bai et al. (2025b) as base model										
Qwen2.5-VL-7B-Instruct	7B	24.4	13.6	6.7	15.9	7.1	4.2	35.3	20.0	12.5
UniTime Li et al. (2026)	7B	–	22.8	14.1	–	41.0	31.5	–	–	–
VideoChat-R1.5 Yan et al. (2026)	7B	52.4	32.3	16.8	71.4	55.8	38.4	–	–	–
Time-R1 Wang et al. (2026)	7B	58.6	39.0	21.4	80.3	66.2	44.8	41.8	29.4	16.4
Ours	**7B**	**65.4**	**44.3**	**25.8**	**86.6**	**73.8**	**51.4**	**60.2**	**42.8**	**25.4**
Qwen2.5-VL-3B-Instruct Bai et al. (2025b) as base model										
Qwen2.5-VL-3B-Instruct	3B	26.5	15.2	7.2	16.8	9.9	3.1	26.1	17.6	10.3
Time-R1 Wang et al. (2026)	3B	46.2	26.0	11.4	40.8	19.7	5.9	40.1	24.4	11.8
Ours	**3B**	**55.2**	**36.8**	**17.4**	**69.3**	**46.1**	**23.9**	**49.3**	**31.2**	**17.5**
Qwen3-VL-8B-Instruct Bai et al. (2025a) as base model										
Qwen3-VL-8B-Instruct	8B	45.9	31.7	19.8	57.2	45.9	34.5	37.3	26.5	15.0
Qwen3-VL-8B-Think Bai et al. (2025a)	8B	44.0	30.5	19.2	69.4	56.3	42.7	38.1	25.9	13.7
Ours	**8B**	**69.7**	**47.6**	**29.1**	**90.3**	**78.2**	**56.5**	**65.1**	**47.4**	**29.8**

datasets (YT-Temporal Zellers et al. (2022), DiDeMo Anne Hendricks et al. (2017), QuerYD Oncescu et al. (2021), InternVid Wang et al. (2024b), HowTo100M Miech et al. (2019)) with temporal frame-shuffling perturbations to refine chronological trajectory generation and temporal grounding accuracy. Besides, we leverage static and multi-image visual reasoning data: Flickr30k LMMS-Lab (2024a), OpenImages Google (2022), CUB Kang (2023) for Stage 1, and GQA LMMS-Lab (2024b), VSR LTL (2023), Visual7W Zhu (2016) fo Stage 2 to enhance complex reasoning tasks.

Our test datasets comprise visual mathematical reasoning (Math-Vista Lu et al. (2023), Math-Vision Wang et al. (2024a)), general multimodal understanding (MMStar Chen et al. (2024), MMMU Yue et al. (2024)), and video temporal grounding and question answering (ActivityNet Krishna et al. (2017), QVHighlights Lei et al. (2021), TVGBench Wang et al. (2026), TACoS Regneri et al. (2013), and Ego4D NLQ Xu et al. (2022)).

4.2 Evaluation Results

As shown in Table 1, our proposed method achieves state-of-the-art (SOTA) performance across all sub-tasks on four mainstream multi-modal benchmarks. Compared to training-free baselines, our method achieves significant performance gains. For instance, on the Qwen2.5-VL-7B Bai et al. (2025b) backbone, ours reaches 80.5% on Math-Vista (OE) and 87.3% on Math-Vista (MC) Lu et al. (2023). Compared to training-based frameworks, our approach demonstrates superior performance, consistently outperforming them across all evaluated dimensions. On MMMU, ours achieves 72.8% on STEM and 80.6% on HASS Yue et al. (2024), significantly outperforming competitive training-based methods like We-Think Yang et al. (2025b) (50.8% and 62.5%, respectively).

Table 3: Performance comparison with existing methods.

Method	TACoS Regneri et al. (2013)			Ego4D NLQ Xu et al. (2022)		
	R1@0.3	R1@0.5	R1@0.7	R1@0.3	R1@0.5	R1@0.7
Qwen2.5-VL-7B-Instruct Bai et al. (2025b) as base model						
VideoChat-R1.5 Yan et al. (2026)	28.6	16.2	6.2	3.8	1.8	0.9
Time-R1 Wang et al. (2026)	36.8	21.5	7.9	6.5	3.3	1.5
Ours	**42.2**	**27.3**	**12.7**	**11.9**	**8.2**	**5.6**
Qwen3-VL-8B-Instruct Bai et al. (2025a) as base model						
Qwen3-VL-8B-Instruct	38.6	30.5	20.4	5.1	2.9	1.5
Ours	**55.8**	**40.5**	**26.3**	**16.7**	**10.2**	**7.8**

Table 4: Ablation study and performance comparison across different stages and configurations on various benchmarks. All models are evaluated based on Pass@1 metric.

Method / Configuration	Math-Vista Lu et al. (2023)		Math-Vision Wang et al. (2024a)		MMStar Chen et al. (2024)		MMMU Yue et al. (2024)	
	OE	MC	OE	MC	PE	RE	STEM	HASS
Base Baseline (Qwen2.5-VL-7B Bai et al. (2025b) Pass@1)	62.3	66.8	24.2	21.7	58.4	68.2	40.1	60.7
A: Stage 1 - Metacognitive Visual Learning (MVL)								
w/o Rcalib (Confidence Calibration)	66.5	73.0	37.1	35.3	60.2	70.9	49.7	65.1
w/o Rsensitivity (Perception Sensitivity)	73.2	79.8	44.6	46.2	66.8	74.1	60.5	71.8
+ Stage 1 Full (MVL) *(Stage 1 End)*	**74.8**	**81.2**	**47.5**	**50.6**	**68.1**	**75.4**	**62.8**	**73.5**
Δ *vs. Base Baseline*	*12.5*	*14.4*	*23.3*	*28.9*	*9.7*	*7.2*	*22.7*	*12.8*
B: Stage 2 - Chronological Trajectory Refinement (CTR)								
w/o Δ Logittemporal (Temporal Perturbation)	76.2	82.5	50.1	53.8	70.2	77.0	65.1	75.8
+ Stage 2 Full (CTR) *(Stage 2 End)*	**77.9**	**84.1**	**53.4**	**57.2**	**71.8**	**78.5**	**68.3**	**77.2**
Δ *vs. Stage 1 Full*	*3.1*	*2.9*	*5.9*	*6.6*	*3.7*	*3.1*	*5.5*	*3.7*
C: Stage 3 - Multi-Horizon Latent Alignment (MHLA)								
w/o Dynamic Gating (η-Threshold, Fixed α)	78.8	85.3	55.6	59.8	73.1	79.8	70.4	78.9
w/o LMHLA (Only Standard CE Loss)	77.9	84.1	53.4	57.2	71.8	78.5	68.3	77.2
CHARM (Full Method) *(Stage 3 End)*	**80.5**	**87.3**	**59.2**	**63.1**	**74.7**	**81.4**	**72.8**	**80.6**
Δ *vs. Stage 2 Full*	*2.6*	*3.2*	*5.8*	*5.9*	*2.9*	*2.9*	*4.5*	*3.4*
Overall Gain	**18.2**	**20.5**	**35.0**	**41.4**	**16.3**	**13.2**	**32.7**	**19.9**

In addition, our method demonstrates strong zero-shot generalization capabilities on video temporal grounding benchmarks (Table 2), outperforming existing state-of-the-art approaches as well as RL-based methods (such as Time-R1 Wang et al. (2026)). Furthermore, our framework consistently yields performance gains across multimodal large language models (MLLMs) of various sizes. Because smaller models possess weaker initial reasoning and grounding capabilities, making it difficult for them to discover optimal paths through random exploration, our framework provides high-quality exemplars and explicit guidance, thereby achieving even more pronounced breakthroughs in smaller models. By grounding its reasoning in key visual cues and specific timestamps, the trained model displays superior grounding performance compared to both the base instruction-tuned model and thinking-enhanced variants.

Our method also presents strong zero-shot performance on long-video datasets (Table 3). When consistently using the Qwen2.5-VL-7B Bai et al. (2025b) backbone, our method achieves SOTA performance, outperforming other baseline methods. Upon upgrading to the newer and stronger Qwen3-VL-8B Bai et al. (2025a) architecture, the training framework still yields substantial performance gains (a 17.2% improvement on TACoS Regneri et al. (2013) R1@0.3 and an 11.6% improvement on Ego4D NLQ Xu et al. (2022) R1@0.3). This demonstrates that our framework is not only effective in long-video scenarios but also generalizes well to stronger, newer model architectures.

To validate the effectiveness of each component, we conduct a comprehensive ablation study across all three progressive stages (Table 4). In Stage 1 (MVL), removing either the confidence calibration reward (R_{calib}) or the perception sensitivity reward ($R_{\text{sensitivity}}$) leads to a noticeable performance drop, confirming that perception and confidence alignment are critical for establishing a solid baseline. Incorporating Stage 2 (CTR) with temporal perturbation penalties ($\Delta\text{Logit}_{\text{temporal}}$) further refines chronological reasoning trajectories, yielding consistent accuracy improvements across all benchmarks. Finally, Stage 3 (MHLA) leverages entropy-guided dynamic gating and latent alignment at test-time; removing dynamic gating (fixed α) or replacing MHLA with standard cross-entropy loss severely impairs performance. The progressive combination of these three stages culminates in substantial overall gains.

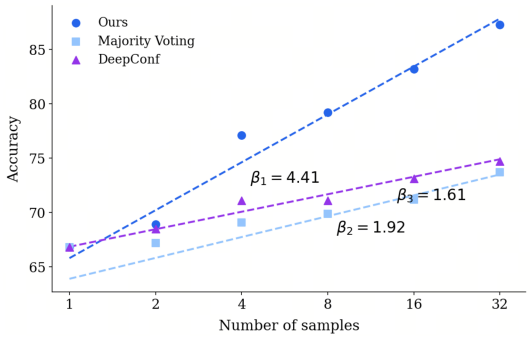
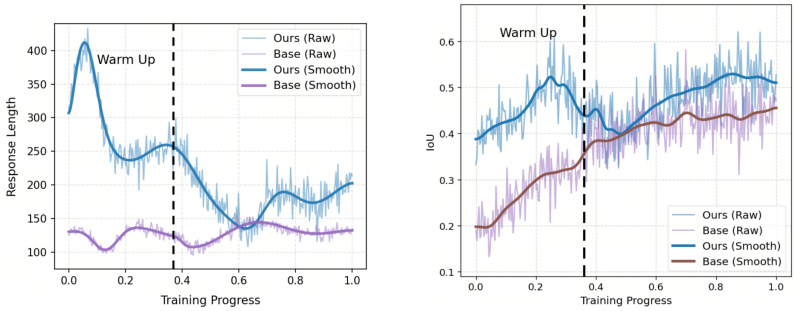


Figure 3: Scaling Efficiency and Performance Comparison (examples from Math-Vista Lu et al. (2023)).



(a) (b)

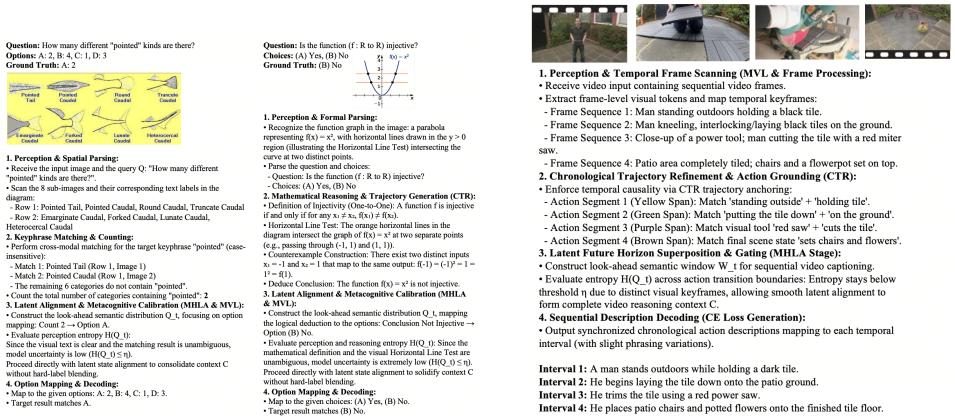
Figure 4: Training dynamics over time: (a) Response length vs. training progress, and (b) IoU vs. training progress.

4.3 Additional Details

Our method significantly improves the model's test-time compute scaling efficiency (Figure 3). Achieving a scaling slope of 4.41, our approach outperforms Majority Voting (1.92) and DeepConf Fu et al. (2025b) (1.61) by a factor of 2.30× to 2.74×. This indicates that CHARM delivers higher returns for the same increase in sampling budget. As the sample size expands from 1 to 32, baseline methods hit a performance bottleneck at around 70% accuracy, whereas CHARM continues to climb up to 87.3%, demonstrating far greater scaling potential.

The response length in the warmup phase of Figure 4(a) reflects the length of our constructed reasoning path; after warmup, a downward trend is shown in the figure. This indicates that the model has learned to filter out irrelevant information rather than verbosely describing every event, thereby ensuring high accuracy without compromising reasoning efficiency. Warmup in Figure 4(b) provides a significantly higher initial reward than the baseline, confirming that the quality of the constructed reasoning path is superior to random rollouts. Notably, we observe a brief drop in IoU after warmup ends, as the model transitions from imitating high-quality external reasoning trajectories to autonomous generation. As Chronological Trajectory Refinement (CTR) continues, IoU steadily rises. This demonstrates that the model successfully discovers better reasoning strategies through self-exploration, proving the necessity of temporal sensitivity and explicit reasoning path generation.

Figure 5 illustrates qualitative execution traces of the proposed framework across static multimodal reasoning benchmarks and dynamic video temporal grounding tasks. For static diagrammatic and mathematical queries, the framework first performs visual perception and formal parsing, leveraging Metacognitive Visual Learning (MVL) and Chronological Trajectory Refinement (CTR) to execute cross-modal keyphrase matching and rigorous logical deduction. Subsequently, Multi-Horizon Latent Alignment (MHLA) quantifies the perception entropy $H(Q_t)$, ensuring calibrated option mapping under low-uncertainty regimes. In dynamic video scenarios, the framework extracts frame-level temporal tokens, anchors causal action trajectories via CTR, and applies MHLA gating to



(a) (b)

Figure 5: (a) Examples from MMStar Chen et al. (2024) and Math-Vista Lu et al. (2023). (b) Examples from ActivityNet Krishna et al. (2017).

smooth look-ahead semantic windows W_t, ultimately producing chronologically coherent video descriptions. These qualitative instances demonstrate the efficacy of our pipeline in integrated visual grounding, calibrated step-wise deduction, and multimodal alignment.

5 Conclusion

In this paper, we propose CHARM, a progressive three-stage framework comprising Metacognitive Visual Learning (MVL), Chronological Trajectory Refinement (CTR), and Multi-Horizon Latent Alignment (MHLA). By combining RL-based confidence calibration, temporal perturbation penalties, and test-time entropy-guided latent alignment, CHARM effectively enhances multimodal reasoning and temporal grounding. Experimental results of Qwen2.5-VL-7B Bai et al. (2025b) show that CHARM achieves state-of-the-art performance across diverse benchmarks, reaching 80.5% (OE) and 87.3% (MC) on Math-Vista Lu et al. (2023), and 72.8% (STEM) and 80.6% (HASS) on MMMU Yue et al. (2024), significantly outperforming competitive training-based baselines like We-Think Yang et al. (2025b) (50.8% and 62.5%). When using Qwen3-VL-8B-Instruct Bai et al. (2025a) as the base model, our method achieves high scores, reaching 69.7 (R1@0.3) on the ActivityNet Krishna et al. (2017) benchmark. Meanwhile, this configuration sets new records on QVHighlights Lei et al. (2021) and TVGBench Wang et al. (2026), achieving 90.3 and 65.1 (R1@0.3).

6 Limitations

First, Stage 3 (MHLA) introduces real-time entropy estimation and dynamic target blending during test-time decoding, which increases per-token inference latency and restricts efficient deployment on resource-constrained or edge devices. Second, key parameters—such as the temporal perturbation step ($\Delta t = 3$s) and the metacognitive entropy threshold ($\eta = 0.6$)—are currently hard-coded static priors rather than being dynamically learned end-to-end, potentially limiting adaptability across video streams with highly varying framerates. Third, despite strong performance on long videos, processing extremely dense visual streams remains constrained by the hard limit of 3,584 video tokens, which can still lead to fine-grained visual information loss in ultra-long, complex video scenarios. In future work, we plan to explore lightweight test-time decoding acceleration, fully adaptive threshold learning, and dynamic sub-sampling token compression to enable real-time, arbitrary-length multimodal reasoning.

References

L. Anne Hendricks, O. Wang, E. Shechtman, J. Sivic, T. Darrell, and B. Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE International Conference on Computer*

Vision, pp. 5803–5812, 2017.

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

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, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, Junyang Lin, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Zhipeng Cai, Ching-Feng Yeh, Hu Xu, Zhuang Liu, Gregory Meyer, Xinjie Lei, Changsheng Zhao, Shang-Wen Li, Vikas Chandra, and Yangyang Shi. Depthlm: Metric depth from vision language models. *arXiv preprint arXiv:2509.25413*, 2025.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? In *Advances in Neural Information Processing Systems*, volume 37, pp. 27056–27087, 2024.

Alberto Compagnoni, Davide Caffagni, Nicholas Moratelli, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. Mitigating hallucinations in multimodal llms via object-aware preference optimization. *arXiv preprint arXiv:2501.00000*, 2025.

Yuetian Du, Yucheng Wang, Ming Kong, Tian Liang, Qiang Long, Bingdi Chen, and Qiang Zhu. Confidence calibration for multimodal llms: An empirical study through medical vqa. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*. Springer Nature Switzerland, 2025.

K. Feng, M. Zhang, H. Li, K. Fan, S. Chen, Y. Jiang, D. Zheng, P. Sun, Y. Zhang, H. Sun, et al. Onethinker: All-in-one reasoning model for image and video. *arXiv preprint arXiv:2512.03043*, 2025.

K. Feng, K. Gong, B. Li, Z. Guo, Y. Wang, T. Peng, J. Wu, X. Zhang, B. Wang, and X. Yue. Video-r1: Reinforcing video reasoning in mllms. In *Advances in Neural Information Processing Systems*, volume 38, pp. 99114–99137, 2026.

Yichao Fu, Xuewei Wang, Yuandong Tian, and Jiawei Zhao. Deep think with confidence. *arXiv preprint arXiv:2508.15260*, 2025a.

Yichao Fu, Xuewei Wang, Yuandong Tian, and Jiawei Zhao. Deep think with confidence. *arXiv preprint arXiv:2508.15260*, 2025b.

J. Geng, F. Cai, Y. Wang, et al. A survey of confidence estimation and calibration in large language models. *arXiv preprint arXiv:2311.08298*, 2023.

Google. Open images dataset v7 facts and figures. `https://storage.googleapis.com/openimages/web/factsfigures_v7.html`, 2022.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*, 2024.

Geoffrey E. Hinton, Simon Osindero, and Yee Whye Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554, 2006.

Jie Huang, Xuejing Liu, Sibo Song, Ruibing Hou, Hong Chang, Junyang Lin, and Shuai Bai. Revisiting multi-modal positional encoding in vision-language models. *arXiv preprint arXiv:2510.23095*, 2025.

Ziwei Ji, Lei Yu, Yeskendir Koishekenov, Yejin Bang, Anthony Hartshorn, Alan Schelten, Cheng Zhang, Pascale Fung, and Nicola Cancedda. Calibrating verbal uncertainty as a linear feature to reduce hallucinations. *arXiv preprint arXiv:2501.07548*, 2025.

Cassie Kang. Cub200 dataset. `https://huggingface.co/datasets/cassiekang/cub200_dataset`, 2023.

R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 706–715, 2017.

J. Lei, T. L. Berg, and M. Bansal. Detecting moments and highlights in videos via natural language queries. In *Advances in Neural Information Processing Systems*, volume 34, pp. 11846–11858, 2021.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. *arXiv preprint arXiv:2311.16922*, 2023.

Bangzheng Li et al. Latent visual reasoning. *arXiv preprint arXiv:2509.24251*, 2025.

Z. Li, S. Di, Z. Zhai, W. Huang, Y. Wang, and W. Xie. Universal video temporal grounding with generative multi-modal large language models. In *Advances in Neural Information Processing Systems*, 2026.

K. Q. Lin, P. Zhang, J. Chen, S. Pramanick, D. Gao, A. J. Wang, R. Yan, and M. Z. Shou. Univtg: Towards unified video-language temporal grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2794–2804, 2023.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2024.

LMMS-Lab. Flickr30k dataset. `https://huggingface.co/datasets/lmms-lab-encoder/flickr30k`, 2024a.

LMMS-Lab. Gqa dataset. `https://huggingface.co/datasets/lmms-lab-encoder/GQA`, 2024b.

Cambridge LTL. Visual spatial reasoning. `https://github.com/cambridgeltl/visual-spatial-reasoning`, 2023.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.

Jizheng Ma, Xiaofei Zhou, Yanlong Song, and Han Yan. Cocova: Chain of continuous vision-language thought for latent space reasoning. *arXiv preprint arXiv:2511.02360*, 2025.

A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, and J. 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*, pp. 2630–2640, 2019.

David Navon. Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology*, 9(3):353–383, 1977.

A.-M. Oncescu, J. F. Henriques, Y. Liu, A. Zisserman, and S. Albanie. Queryd: A video dataset with high-quality text and audio narrations. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2265–2269. IEEE, 2021.

OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

J. Park, J. Na, J. Kim, and H. J. Kim. Deepvideo-r1: Video reinforcement fine-tuning via difficulty-aware regressive grpo. In *Advances in Neural Information Processing Systems*, volume 38, pp. 138605–138632, 2026.

M. Regneri, M. Rohrbach, D. Wetzel, S. Thater, B. Schiele, and M. Pinkal. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1:25–36, 2013.

S. Ren, L. Yao, S. Li, X. Sun, and L. Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14313–14323, 2024.

Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. Li, Y. Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Xing Shen, Justin Szeto, Mingyang Li, Hengguan Huang, and Tal Arbel. Exposing and mitigating calibration biases and demographic unfairness in mllm few-shot in-context learning for medical image classification. *arXiv preprint arXiv:2501.00000*, 2025.

Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *arXiv preprint arXiv:2503.00000*, 2025a.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. In *Advances in Neural Information Processing Systems*, volume 37, pp. 95095–95169, 2024a.

Qixun Wang, Yang Shi, Yifei Wang, Yuanxing Zhang, Pengfei Wan, Kun Gai, Xianghua Ying, and Yisen Wang. Monet: Reasoning in latent visual space beyond images and language. *arXiv preprint arXiv:2511.21395*, 2025b.

Y. Wang, Y. He, Y. Li, K. Li, J. Yu, X. Ma, X. Li, G. Chen, X. Chen, Y. Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. In *International Conference on Learning Representations*, volume 2024, pp. 42055–42079, 2024b.

Y. Wang, Z. Wang, B. Xu, Y. Du, K. Lin, Z. Xiao, Z. Yue, J. Ju, L. Zhang, D. Yang, et al. Time-r1: Post-training large vision language model for temporal video grounding. In *Advances in Neural Information Processing Systems*, volume 38, pp. 83330–83364, 2026.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V. Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022.

xAI Organization. Realworldqa dataset. `https://huggingface.co/datasets/xai-org/RealworldQA`, 2024.

Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts. *arXiv preprint arXiv:2407.04973*, 2024.

M. Xu, C. Zhao, M. Ramazanova, and B. Ghanem. Ego4d: Around the world in 3,000 hours of egocentric video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18973–18990. IEEE, 2022.

Z. Yan, Y. He, X. Li, Z. Yue, X. Zeng, Y. Wang, Y. Qiao, L. Wang, and Y. Wang. Videochat-r1.5: Visual test-time scaling to reinforce multimodal reasoning by iterative perception. In *Advances in Neural Information Processing Systems*, 2026.

J. Yang, M. Ma, Y. Li, and Z. Pei. Vqals: A video question answering method in low-light scenes based on illumination correction and feature enhancement. *Chinese Journal of Electronics*, 34(4): 1300–1308, 2025a.

Jie Yang, Feipeng Ma, Zitian Wang, Dacheng Yin, Kang Rong, Fengyun Rao, and Ruimao Zhang. Wethink: Toward general-purpose vision-language reasoning via reinforcement learning. *arXiv preprint arXiv:2506.07905*, 2025b.

Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, et al. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*, 2025c.

Shaofeng Yin, Ting Lei, and Yang Liu. Toolvqa: A dataset for multi-step reasoning vqa with external tools. *arXiv preprint arXiv:2508.03284*, 2025.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*, 2025.

Eric Zelikman, Georges Raif Harik, Yijia Shao, Varuna Jayasiri, Nick Haber, and Noah Goodman. Quiet-star: Language models can teach themselves to think before speaking. In *First Conference on Language Modeling*, 2024.

R. Zellers, J. Lu, X. Lu, Y. Yu, Y. Zhao, M. Salehi, A. Kusupati, J. Hessel, A. Farhadi, and Y. Choi. Merlot reserve: Neural script knowledge through vision and language and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16375–16387, 2022.

Huanyu Zhang, Wenshan Wu, Chengzu Li, Ning Shang, Yan Xia, Yangyu Huang, Yifan Zhang, Li Dong, Zhang Zhang, Liang Wang, et al. Latent sketchpad: Sketching visual thoughts to elicit multimodal reasoning in mllms. *arXiv preprint arXiv:2510.24514*, 2025.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, and Hongsheng Li. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? *arXiv preprint arXiv:2403.14624*, 2024.

Yuke Zhu. Visual7w toolkit. `https://github.com/yukezhu/visual7w-toolkit`, 2016.

Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. Dynamath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models. *arXiv preprint arXiv:2411.00835*, 2025.

APPENDIX A ALGORITHMS

Algorithm 1 Stage 1: Metacognitive Visual Learning (MVL)

Require: Initial model parameters θ_0; Training dataset $\mathcal{D}_{\text{MVL}} = \{(I, Q, y)\}$; Calibration scaling hyperparameter λ_{calib}; Sensitivity scaling hyperparameter λ_{sens}.
Ensure: Calibrated base model parameters θ_{MVL}.
1: Initialize model parameters $\theta \leftarrow \theta_0$
2: **while** not converged **do**
3: Sample a mini-batch $\{(I, Q, y)\} \sim \mathcal{D}_{\text{MVL}}$
4: Apply visual/perception perturbation to input I to construct perturbed context I'
5: Compute task accuracy reward $\mathcal{R}_{\text{acc}}$ evaluating prediction correctness under I
6: Compute confidence calibration reward $\mathcal{R}_{\text{calib}}$ penalizing misjudgments stemming from over-confidence or under-confidence
7: Compute perception sensitivity reward $\mathcal{R}_{\text{sensitivity}}$ enforcing a drop in confidence under perturbed context I'
8: Formulate joint MVL reward: $\mathcal{R}_{\text{MVL}} \leftarrow \mathcal{R}_{\text{acc}} + \lambda_{\text{calib}} \cdot \mathcal{R}_{\text{calib}} + \lambda_{\text{sens}} \cdot \mathcal{R}_{\text{sensitivity}}$
9: Update parameters θ via Group Relative Policy Optimization (GRPO) using objective $\mathcal{R}_{\text{MVL}}$
10: **end while**
11: **return** $\theta_{\text{MVL}} \leftarrow \theta$

Algorithm 2 Stage 2: Chronological Trajectory Refinement (CTR)

Require: Calibrated model parameters θ_{MVL}; Sequential training dataset $\mathcal{D}_{\text{CTR}} = \{(I, Q, y)\}$; Temporal perturbation weighting hyperparameter γ.
Ensure: Temporally refined model parameters θ_{CTR}.
1: Initialize model parameters $\theta \leftarrow \theta_{\text{MVL}}$
2: **while** not converged **do**
3: Sample a mini-batch $\{(I, Q, y)\} \sim \mathcal{D}_{\text{CTR}}$
4: Perform frame shuffling on video context I to yield temporally perturbed input I_{shuffled}
5: Compute temporal grounding/accuracy reward $\mathcal{R}_{\text{IoU/Acc}}$ evaluating spatial-temporal keyframe localization
6: Measure prediction logit variance under temporal frame shuffling: $\Delta\text{Logit}_{\text{temporal}} \leftarrow \text{Logit}(I) - \text{Logit}(I_{\text{shuffled}})$
7: Compute total CTR reward: $\mathcal{R}_{\text{CTR}} \leftarrow \mathcal{R}_{\text{IoU/Acc}} + \gamma \cdot \Delta\text{Logit}_{\text{temporal}}$
8: Update parameters θ via Reinforcement Learning driven by $\mathcal{R}_{\text{CTR}}$ to anchor trajectories to keyframes
9: **end while**
10: **return** $\theta_{\text{CTR}} \leftarrow \theta$

Algorithm 3 Stage 3: Multi-Horizon Latent Alignment (MHLA)

Require: Temporally refined model parameters θ_{CTR}; Input visual context I and query Q; Ground-truth answer token sequence $y = \{a_1, a_2, \ldots, a_M\}$; Maximum latent reasoning steps T; Entropy gating threshold η; Blending hyperparameter α.
Ensure: Optimized parameters θ^* and generated textual answer sequence $A = \{a_1, a_2, \ldots, a_M\}$.
Initialize model parameters $\theta \leftarrow \theta_{\text{CTR}}$, latent context $\mathcal{C} \leftarrow \emptyset$
// Phase I: Multi-Horizon Latent Reasoning
for $t = 1 \rightarrow T$ **do**
  Identify dynamic look-ahead semantic window W_t over ungenerated future tokens
  Construct target superposition distribution: $Q_t \leftarrow \sum_{k \in W_t} w_k \cdot P_\theta(\cdot \mid I, Q, c_{<t}, k)$
  Compute Shannon entropy for uncertainty estimation: $H(Q_t) \leftarrow -\sum Q_t \log Q_t$
  if $H(Q_t) > \eta$ **then**
    Construct uncertainty-gated target distribution: $P_t^{\text{target}} \leftarrow \alpha \cdot y_{\text{hard}} + (1 - \alpha) \cdot Q_t$
  else
    Set soft target distribution: $P_t^{\text{target}} \leftarrow Q_t$
  end if
  Compute Multi-Horizon Latent Alignment Loss:
    $\mathcal{L}_{\text{MHLA}} \leftarrow -\frac{1}{T} \sum_{t=1}^{T} \sum_{k \in W_t} P_t^{\text{target}}(k) \log P_\theta(k \mid I, Q, c_{<t})$
  Update latent step state c_t and accumulate context: $\mathcal{C} \leftarrow \mathcal{C} \cup \{c_t\}$
end for
// Phase II: Explicit Textual Answer Decoding
Compute explicit answer Cross-Entropy Loss: $\mathcal{L}_{\text{CE}} \leftarrow -\frac{1}{M} \sum_{j=1}^{M} \log P_\theta(a_j \mid I, Q, \mathcal{C}, a_{<j})$
// Joint Optimization & Generation
Formulate combined loss: $\mathcal{L}_{\text{Total}} \leftarrow \mathcal{L}_{\text{MHLA}} + \mathcal{L}_{\text{CE}}$
Optimize model parameters θ under $\mathcal{L}_{\text{Total}}$ to obtain θ^*
Autoregressively generate explicit sequence $A = \{a_1, a_2, \ldots, a_M\}$ conditioned on accumulated latent context $\mathcal{C}$
return θ^*, A

APPENDIX B PROMPTS

To provide structured supervision during progressive reinforcement learning, stage-specific system prompts are engineered for Stage 1 (MVL) and Stage 2 (CTR) (Figure 6). In Stage 1, the prompt is tailored to cultivate metacognitive self-calibration and perception sensitivity across general visual question answering and visual-mathematical reasoning tasks. It explicitly instructs the model to extract fine-grained visual-logical evidence, self-evaluate internal prediction certainty through a calibrated confidence score, and analyze susceptibility to visual and logical perturbations, thereby establishing reliable uncertainty estimates. In Stage 2, the prompt transitions toward enforcing spatial-temporal grounding and chronological trajectory generation. Specifically optimized for video sequence and multi-image relational datasets, it guides the model to anchor its reasoning to precise keyframe timestamps or spatial-relational attributes, while constructing explicit trajectories that track event progressions in strict chronological order. Together, these prompts standardize the reasoning space, providing transparent intermediate targets that directly facilitate joint reward formulation across both alignment stages.

APPENDIX C ABLATION STUDIES

To comprehensively evaluate the effectiveness of each component in our proposed CHARM framework, we conduct progressive ablation experiments (Table 5 and Table 6) on Qwen3-VL-2B-Think Bai et al. (2025a) and Qwen2.5-VL-7B-Instruct Bai et al. (2025b) backbones across a wide spectrum of benchmarks, including mathematical reasoning, multimodal understanding, and temporal video grounding. As demonstrated in the results, Stage 1 (MVL) significantly enhances baseline metacognitive and perceptual capabilities through confidence calibration and perception sensitivity. Building upon this, Stage 2 (CTR) incorporates temporal perturbation mitigation to further refine reasoning trajectories. Finally, Stage 3 (MHLA) achieves multi-horizon latent alignment via dynamic gat-

Prompt for Stage 1: Metacognitive Visual Learning (MVL)

You are an advanced multimodal metacognitive reasoner skilled in general visual question answering and visual mathematical/logical reasoning.

Your task is to answer the query while executing a rigorous metacognitive self-calibration:
1. Visual & Mathematical Observation: Extract key visual elements, spatial details, and numerical/mathematical constraints from the input.
2. Metacognitive Calibration: Assess your internal certainty based on visual clarity and reasoning logic. Explicitly rate your Confidence Score (0.00 to 1.00) and identify potential sensitivity to visual/logic perturbations.
3. Step-by-Step Reasoning: Provide a logical or mathematical derivation grounded in the visual evidence.
4. Final Answer: Provide the precise final answer.

Output Structure:
- [Visual & Logical Evidence]: <Observations and extracted mathematical/logical conditions>
- [Metacognitive Assessment]:
  * Perception/Logic Sensitivity: High / Medium / Low
  * Internal Confidence: <Score between 0.00 and 1.00>
  * Calibration Rationale: <Why this score was assigned>
- [Reasoning Path]: <Step-by-step visual derivation or math solution>
- [Final Answer]: <Concise final response>

Prompt for Stage 2: Chronological Trajectory Refinement (CTR)

You are a spatial-temporal reasoning expert specializing in video keyframe trajectory tracking and spatial-relational grounding.

Your task is to anchor your reasoning strictly to temporal keyframes (for videos) or spatial relations (for multi-image/static scenes):
1. Spatial-Temporal Grounding:
  - For Videos: Identify the critical timestamps or keyframe indices.
  - For Images: Identify spatial bounding regions, object relationships, or directional attributes.
2. Trajectory Refinement: Track the chronological progression of events or logical spatial dependencies without temporal/spatial hallucination.
3. Perturbation Robustness: Ensure the reasoning directly depends on exact chronological sequence and spatial layout.

Output Structure:
- [Spatial-Temporal Grounding]: <Keyframe timestamps [t_start, t_end] OR spatial relation tags (e.g., left_of, inside)>
- [Refined Trajectory]:
  * Step 1: Initial state / spatial baseline.
  * Step 2: Temporal progression / spatial interaction.
  * Step 3: Final state / target condition.
- [Final Answer]: <Concise answer or grounded timestamp/location>

Figure 6: Prompts for Stage 1 and Stage 2.

Table 5: Ablation Study based on Qwen3-VL-2B-Think Bai et al. (2025a) model.

Method / Configuration	Math-Vista Lu et al. (2023)		Math-Vision Wang et al. (2024a)		MMStar Chen et al. (2024)		MMMU Yue et al. (2024)	
	OE	MC	OE	MC	PE	RE	STEM	HASS
Qwen3-VL-2B-Think Bai et al. (2025a) Pass@1	65	82.5	47.7	50	62.8	56	58.8	67
A: Stage 1 - Metacognitive Visual Learning (MVL)								
w/o Rcalib (Confidence Calibration)	69.2	85.1	52.1	55.3	65.8	62.1	64.5	72.2
w/o Rsensitivity (Perception Sensitivity)	73.5	88	55.6	60.2	67.9	68	70.6	76.5
+ Stage 1 Full (MVL) *(Stage 1 End)*	**74.8**	**88.8**	**57.2**	**61.5**	**69.8**	**69.5**	**71.8**	**78.2**
Δ *vs. Base Baseline*	*9.8*	*6.3*	*9.5*	*11.5*	*7*	*13.5*	*13*	*11.2*
B: Stage 2 - Chronological Trajectory Refinement (CTR)								
w/o Δ Logittemporal (Temporal Perturbation)	76.5	89.8	59.5	64.2	71.5	73.2	75	81.1
+ Stage 2 Full (CTR) *(Stage 2 End)*	**78.4**	**90.9**	**61.8**	**66.5**	**72.9**	**75.8**	**77.4**	**83.2**
Δ *vs. Stage 1 Full*	*3.6*	*2.1*	*4.6*	*5*	*3.1*	*6.3*	*5.6*	*5*
C: Stage 3 - Multi-Horizon Latent Alignment (MHLA)								
w/o Dynamic Gating (η-Threshold, Fixed α)	80.8	92.2	63.6	68.8	74.1	78.4	79.7	85.9
w/o LMHLA (Only Standard CE Loss)	78.2	90.8	61.9	66.7	72.6	75.7	77.2	83.2
CHARM (Full Method) *(Stage 3 End)*	**82.9**	**93.7**	**65.4**	**71.2**	**75.5**	**81.3**	**82.5**	**88.1**
Δ *vs. Stage 2 Full*	*4.5*	*2.8*	*3.6*	*4.7*	*2.6*	*5.5*	*5.1*	*4.9*
Overall Gain	**17.9**	**11.2**	**17.7**	**21.2**	**12.7**	**25.3**	**23.7**	**21.1**

ing. The three stages operate synergistically, enabling the full CHARM framework to consistently achieve optimal performance across all datasets.

Besides, we analyze the impact of key hyper-parameters across the progressive optimization and decoding stages (Table 7). For the shuffle window size (Δt) in Stage 2, setting $\Delta t = 3$s achieves peak performance across all benchmarks. A smaller window ($\Delta t = 1$s) provides insufficient temporal perturbation signals, while a larger window ($\Delta t = 5$s) disrupts long-range chronological continuity. Regarding the entropy threshold (η) in Stage 3, $\eta = 0.6$ yields optimal results. Lowering the threshold ($\eta = 0.3$) prematurely triggers hard label substitution for moderately uncertain tokens, whereas setting $\eta = 1.0$ effectively disables the adaptive target switching mechanism, resulting in a noticeable performance drop across all metrics. Finally, for the blending factor (α), a moderate value of 0.8 strikes the optimal trade-off between soft latent alignment and hard label correction. Deviating to lower or higher blending weights leads to suboptimal performance, confirming the necessity of balanced target superposition during test-time decoding.

APPENDIX D EXPERIMENTAL ANALYSIS

Figure 7 compares the accuracy of our method with the Majority Voting and DeepConf Fu et al. (2025b) baselines as the number of samples increases. The evaluation results on MMStar Chen et al. (2024) and MMMU Yue et al. (2024) demonstrate that the slope of our method's trend line is significantly steeper than those of the baselines, indicating that the performance advantage and potential of CHARM become more pronounced with an increased number of samples.

Table 6: Ablation Study based on Qwen2.5-VL-7B-Instruct Bai et al. (2025b) model.

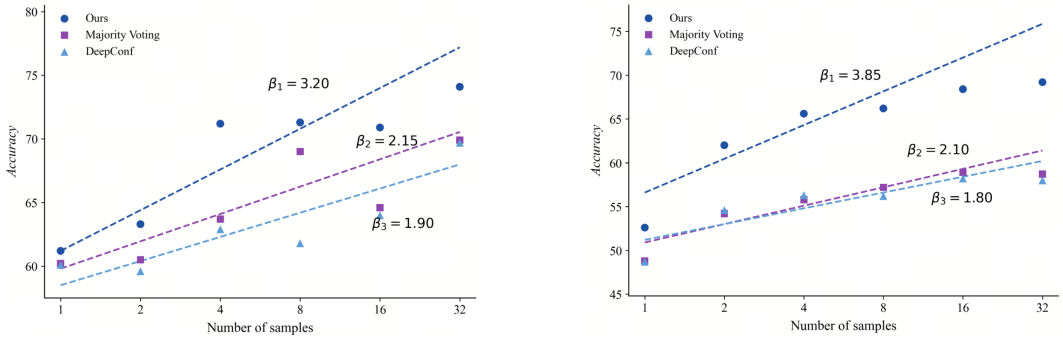
Method / Configuration	ActivityNet Krishna et al. (2017)			QVHighlights Lei et al. (2021)			TVGBench Wang et al. (2026)		
	R1@0.3	R1@0.5	R1@0.7	R1@0.3	R1@0.5	R1@0.7	R1@0.3	R1@0.5	R1@0.7
Qwen2.5-VL-7B-Instruct Bai et al. (2025b)	24.4	13.6	6.7	15.9	7.1	4.2	35.3	20	12.5
A: Stage 1 - Metacognitive Visual Learning (MVL)									
w/o Rcalib (Confidence Calibration)	36.8	21.5	11.8	40.2	24.5	15.9	43.1	27.8	16.8
w/o Rsensitivity (Perception Sensitivity)	38.2	23.1	13.2	43.5	27.8	18.4	45.6	29.5	18.1
+ Stage 1 Full (MVL) *(Stage 1 End)*	**39.5**	**24.8**	**14.5**	**45.8**	**30.2**	**20.5**	**47.2**	**31**	**19.2**
Δ *vs. Base Baseline*	*15.1*	*11.2*	*7.8*	*29.9*	*23.1*	*16.3*	*11.9*	*11*	*6.7*
B: Stage 2 - Chronological Trajectory Refinement (CTR)									
w/o Δ Logittemporal (Temporal Perturbation)	56.5	35.2	19.8	68.4	52.1	36.5	53.8	36.5	21.5
+ Stage 2 Full (CTR) *(Stage 2 End)*	**58.2**	**37.5**	**21.4**	**71.5**	**55.8**	**39.2**	**55.4**	**38.2**	**22.8**
Δ *vs. Stage 1 Full*	*18.7*	*12.7*	*6.9*	*25.7*	*25.6*	*18.7*	*8.2*	*7.2*	*3.6*
C: Stage 3 - Multi-Horizon Latent Alignment (MHLA)									
w/o Dynamic Gating (η-Threshold, Fixed α)	64.2	42.9	25.2	84.8	71.5	49.8	59.1	41.8	24.8
w/o LMHLA (Only Standard CE Loss)	63.1	42.5	24.5	82.5	69.8	48.5	58.2	40.5	24.2
CHARM (Full Method) *(Stage 3 End)*	**65.4**	**44.3**	**25.8**	**86.6**	**73.8**	**51.4**	**60.2**	**42.8**	**25.4**
Δ *vs. Stage 2 Full*	*7.2*	*6.8*	*4.4*	*15.1*	*18*	*12.2*	*4.8*	*4.6*	*2.6*
Overall Gain	**41**	**30.7**	**19.1**	**70.7**	**66.7**	**47.2**	**24.9**	**22.8**	**12.9**

Table 7: Ablation study of hyper-parameters based on Qwen2.5-VL-7B Bai et al. (2025b).

Setting	ActivityNet Krishna et al. (2017)			QVHighlights Lei et al. (2021)			TVGBench Wang et al. (2026)		
	R1@0.3	R1@0.5	R1@0.7	R1@0.3	R1@0.5	R1@0.7	R1@0.3	R1@0.5	R1@0.7
Stage 2: Shuffle Window Size (Δt)									
$\Delta t = 1$s	63.8	42.1	23.9	84.2	70.5	48.2	57.6	39.8	23.1
$\Delta t = 3$s (Default)	**65.4**	**44.3**	**25.8**	**86.6**	**73.8**	**51.4**	**60.2**	**42.8**	**25.4**
$\Delta t = 5$s	62.9	41.5	23.2	83.5	69.8	47.1	56.1	38.2	22.0
Stage 3: Entropy Threshold (η)									
$\eta = 0.3$	61.2	39.8	22.1	81.9	68.2	45.3	55.0	37.4	21.2
$\eta = 0.6$ (Default)	**65.4**	**44.3**	**25.8**	**86.6**	**73.8**	**51.4**	**60.2**	**42.8**	**25.4**
$\eta = 1.0$	62.5	41.2	23.4	83.0	69.4	46.8	56.8	39.1	22.8
Stage 3: Hard Label Blend Factor (α)									
$\alpha = 0.5$	63.1	42.0	23.8	84.1	70.8	48.0	57.2	39.5	23.0
$\alpha = 0.8$ (Default)	**65.4**	**44.3**	**25.8**	**86.6**	**73.8**	**51.4**	**60.2**	**42.8**	**25.4**
$\alpha = 1.0$	63.8	42.6	24.1	84.8	71.3	48.6	58.0	40.2	23.6

APPENDIX E QUALITATIVE CASE ANALYSIS

Figure 8 and 9 illustrate qualitative execution examples of our framework across symbolic mathematical reasoning and temporal video grounding. In the cross-puzzle case, the perception module parses the 5-square spatial layout, after which Calibrated Trajectory Refinement (CTR) evaluates candidate center values $x \in \{1, 2, 3, 4, 9\}$ to uniquely deduce $x = 9$ via the parity constraint $2S_h = 19 + x$. For the calculus task, CTR calculates curve intersection boundaries ($x = 0, 1.5, 4$) between $f(x) = x^3 - 6x^2 + 8x$ and $g(x) = -\frac{1}{2}x^2 + 2x$, directly formulating the target area integral $\int_0^{1.5}[f(x) - g(x)]dx \rightarrow$ Option (A). In both mathematical tasks, Multi-Horizon Latent Alignment (MHLA) verifies low perception entropy ($H(Q_t) \leq \eta$), confirming deterministic latent alignment. Conversely, in video localization for the query *"A few cars in a parking lot"*, the framework scans keyframe tokens, anchors action spans via CTR, and utilizes MHLA look-ahead windows W_t to precisely isolate matching temporal intervals (Intervals 3 and 6). These specific instances demonstrate how our framework unifies discrete algebraic deduction and continuous temporal grounding under metacognitively calibrated uncertainty gating.



(a) MMStar (b) MMMU

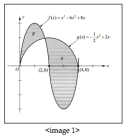
Figure 7: Accuracy comparison on (a) MMStar Chen et al. (2024) and (b) MMMU Yue et al. (2024) as the number of samples increases.

Question: The numbers 1,2,3,4 and 9 are written into the squares on the following figure. The sum of the three numbers in the horizontal row, should be the same as the sum of the three numbers in the vertical column. Which number is written in the middle?
Ground Truth: 9



1. Perception & Spatial Parsing:
Cross/plus-shaped arrangement comprising 5 identical squares (1 center square shared by both the horizontal row of 3 squares and vertical column of 3 squares).
Given number set: {1,2,3,4,9}.
2. Mathematical Analysis & Constraint Solving:
• Formulate algebraic relations for the row and column sums:
Let x denote the middle (center) square's number. Sum of all 5 numbers: $S = 1 + 2 + 3 + 4 + 9 = 19$. Let S_h be the sum of the horizontal row and S_v be the sum of the vertical column ($S_h = S_v$). Since the central square x is included in both the row and column: $S_h + S_v = S + x \Rightarrow 2S_h = 19 + x$.
• Parity and feasibility check for candidate middle numbers $x \in \{1,2,3,4,9\}$:Since $2S_h$ is an even integer, $19 + x$ must be even, restricting x to odd candidates {1,3,9}.
Candidate $x = 1$: $2S_h = 20 \Rightarrow S_h = 10$. Remaining set {2,3,4,9} cannot be partitioned into two pairs each adding to $10 - 1 = 9$. Candidate $x = 3$: $2S_h = 22 \Rightarrow S_h = 11$. Remaining set {1,2,4,9} cannot be partitioned into two pairs each adding to $11 - 3 = 8$. Candidate $x = 9$: $2S_h = 28 \Rightarrow S_h = 14$. Remaining set {1,2,3,4} partitions into pairs (1,4) and (2,3), both adding to $14 - 9 = 5$.
• Determine the unique valid middle value: $x = 9$.
3. Latent Alignment & Metacognitive Calibration (MHLA & MVL):
• Construct the look-ahead semantic distribution Q_t, focusing on numerical mapping: Middle value →9.
• Evaluate perception entropy $H(Q_t)$:
Since the diagram layout is clear and the arithmetic constraint yields an unambiguous unique solution, model uncertainty is low ($H(Q_t) \leq \eta$). Proceed directly with latent state alignment to consolidate context C without hard-label blending.
4. Option Mapping & Decoding: Target result: **9**

Question: <image 1>The region bounded by the graph as shown above. Choose an integral expression that can be used to find the area of R.
(A) $\int_0^{1.5}[f(x) - g(x)]\,dx$
(B) $\int_0^{1.5}[g(x) - f(x)]\,dx$
(C) $\int_0^{2}[f(x) - g(x)]\,dx$
(D) $\int_0^{2}[g(x) - f(x)]\,dx$
Ground Truth: A



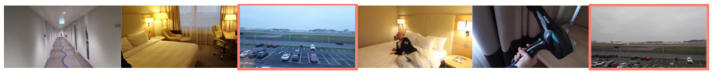
1. Mathematical Reasoning & Trajectory Generation (CTR):
• Integrand Determination: The area bounded between two continuous functions: $f(x) \geq g(x)$ is given by $\int_a^b [f(x) - g(x)]\,dx$. Observing region R, $f(x)$ is the upper curve and $g(x)$ is the lower curve, making the integrand $[f(x) - g(x)]$.
• Intersection & Boundary Calculation: Find the integration bounds by setting $f(x) = g(x)$:

$$x^3 - 6x^2 + 8x = -\frac{1}{2}x^2 + 2x \Rightarrow x^3 - \frac{11}{2}x^2 + 6x = 0$$

Factoring gives $x\left(2x^2 - 11x + 12\right) = x(2x - 3)(x - 4) = 0$, yielding roots $x = 0$, $x = 1.5$, and $x = 4$.
• Area Expression Formulation: Region R spans from $x = 0$ to the first non-zero intersection point at $x = 1.5$. Thus, the integral representing the area of R is $\int_0^{1.5}[f(x) - g(x)]\,dx$.
• Deduce Conclusion: The correct integral expression is $\int_0^{1.5}[f(x) - g(x)]\,dx$.
2. Latent Alignment & Metacognitive Calibration (MHLA & MVL):
• Construct the look-ahead semantic distribution Q_t, mapping the logical deduction to the options: Expression $\int_0^{1.5}[f(x) - g(x)]\,dx \rightarrow$ Option (A).
• Evaluate perception and reasoning entropy $H(Q_t)$: Since the relative spatial orientation of the curves and the algebraic solution for the intersection point $x = 1.5$ are exact and unambiguous, model uncertainty is extremely low ($H(Q_t) \leq \eta$). Proceed directly with latent state alignment to solidify context C without hard-label blending.
3. Option Mapping & Decoding:
• Target result matches (A) $\int_0^{1.5}[f(x) - g(x)]\,dx$.

Figure 8: Math-Vision Wang et al. (2024a) and MMMU Yue et al. (2024) examples.

Question: A few cars are in a parking lot.



1. Perception & Temporal Frame Scanning (MVL & Frame Processing):
• Receive video input containing sequential video frames accompanied by the query: "A few cars are in a parking lot."
• Extract frame-level visual tokens and map temporal keyframes across the sequence:
Frame Sequence 1: Indoor hotel hallway with a patterned carpet.
Frame Sequence 2: Well-lit hotel bedroom interior featuring a made bed and a desk.
Frame Sequence 3 (Highlighted with red border): Outdoor view through a window showing a parking lot with several parked cars.
Frame Sequence 4: Man reclining on a hotel bed.
Frame Sequence 5: Close-up of a handheld electronic device or appliance.
Frame Sequence 6 (Highlighted with red border): Another outdoor view showing a parking lot with cars.

2. Chronological Trajectory Refinement & Action Grounding (CTR):
• Enforce temporal causality via CTR trajectory anchoring to match the target query "A few cars are in a parking lot":
Action Segment 1: Indoor hallway traversal.
Action Segment 2: Hotel bedroom view.
Action Segment 3 (Red Box Span): Matches the query description 'cars in a parking lot'.
Action Segment 4: Person resting on a bed.
Action Segment 5: Close-up handle view.
Action Segment 6 (Red Box Span): Matches the query description 'cars in a parking lot'.

3. Latent Future Horizon Superposition & Gating (MHLA Stage):
• Construct look-ahead semantic window W_t for temporal grounding and localization.
• Evaluate entropy $H(Q_t)$ across frame boundaries: Entropy remains low where visual scenes directly correspond to the descriptive query text, allowing smooth latent alignment to isolate matching intervals.

4. Sequential Description Decoding (CE Loss Generation):
• Output synchronized temporal localization mapping to the matching intervals where cars in a parking lot are visible:
Interval 3: A view through the window shows a few cars in a parking lot.
Interval 6: Another view shows a few cars in a parking lot.

Figure 9: QVHighlights Lei et al. (2021) example.