Training-free Uncertainty Guidance for Complex Visual Tasks with MLLMs

Sanghwan Kim[1,2,3] Rui Xiao[1,2] Stephan Alaniz[5]
Yongqin Xian[4] Zeynep Akata[1,2,3]

[1]Technical University of Munich [2]Munich Center for Machine Learning
[3]Helmholtz Munich [4]Google [5]LTCI, Télécom Paris, Institut Polytechnique de Paris

Abstract. Multimodal Large Language Models (MLLMs) often struggle with fine-grained perception, such as identifying small objects in high-resolution images or detecting key moments in long videos. Existing methods typically rely on complex, task-specific fine-tuning, which reduces generalizability and increases system complexity. In this work, we propose an effective, training-free framework that uses an MLLM's intrinsic uncertainty as proactive guidance. Our core insight is that a model's uncertainty decreases when provided with relevant visual information. We introduce a unified mechanism that scores candidate visual inputs by response uncertainty, enabling the model to autonomously focus on the most informative data. We apply this simple principle to three challenging visual tasks: Visual Search, Long Video Understanding, and Temporal Grounding, allowing off-the-shelf MLLMs to achieve performance competitive with specialized, fine-tuned systems. Our results demonstrate that leveraging intrinsic uncertainty is a powerful strategy for improving fine-grained multimodal performance. Code is available at `https://github.com/ExplainableML/ug-framework`.

Keywords: Multimodal Large Language Models · Fine-Grained Perception · Uncertainty Estimation

1 Introduction

Multimodal Large Language Models (MLLMs) have achieved notable progress in general visual understanding, yet they often underperform on tasks requiring fine-grained or localized perception [39, 40, 54]. A central challenge lies in identifying sparse but crucial information within large and noisy visual contexts. This issue is particularly evident in tasks such as *Visual Search*, which requires detecting small objects in high-resolution images; *Long Video Understanding*, which depends on locating key moments across long video sequences; and *Temporal Grounding*, which involves determining the precise temporal span of an event. In all cases, the vast visual input makes uniform processing impractical, causing models to overlook important details and produce incorrect outputs.

To address this, existing approaches often adopt complex and task-specific solutions. Many involve significant engineering effort, including curating custom

datasets for each task and fine-tuning models, or integrating external, specialized models to select relevant regions [11,19,44,49,52]. Although sometimes effective, their high engineering costs and specificity limit generalizability, making them difficult to adapt across diverse tasks requiring nuanced visual reasoning. The core challenge remains: finding a simple, unified, and training-free strategy to guide MLLMs' focus toward the most relevant visual information.

In this work, we propose a principled alternative based on an MLLM's intrinsic uncertainty. While prior work shows that output uncertainty, often measured by entropy, correlates with hallucinations and is useful for post-hoc error detection [14,34,42], its use as a proactive guidance signal has been largely overlooked. We argue that intrinsic uncertainty provides a direct, real-time indicator of which visual input is most informative. Our key insight is that when an MLLM is shown the correct visual evidence for a query, its predictive confidence increases and uncertainty naturally decreases.

Building on this observation, we introduce the *Uncertainty-Guided (UG) framework*, which reformulates diverse localization tasks as a unified problem: finding the state of minimum uncertainty. The framework evaluates candidate visual inputs (e.g., image crops or video frames) and scores them using the MLLM's response uncertainty. This can be measured via the output distribution's *entropy* or, for binary decisions, a direct *confidence score* from "yes/no" probabilities. This simple mechanism enables the model to autonomously identify and attend to the most informative content. Whether selecting the best crop for visual search, sampling the top-k informative frames for video QA, or locating the highest-confidence segment for temporal grounding, the underlying principle remains consistent.

This unified, training-free framework enables off-the-shelf MLLMs to solve complex localization tasks without architectural changes or fine-tuning. We apply it to multiple open-source models, including the LLaVA [18], Qwen-VL [2], and InternVL [6] families. Our contributions are as follows: (1) We show that minimizing an MLLM's output uncertainty reliably identifies task-relevant visual evidence, establishing uncertainty as a principled localization signal. (2) We introduce a simple, unified, training-free framework that leverages intrinsic uncertainty for diverse fine-grained visual localization tasks without task-specific designs. (3) Extensive experiments demonstrate that our framework enables standard MLLMs to achieve competitive results with state-of-the-art (SOTA) specialized methods on Visual Search, Long Video Understanding, and Temporal Grounding.

2 Related Work

Multimodal Large Language Models (MLLMs). MLLMs have rapidly advanced into powerful systems combining perception and reasoning. Foundational open-source models such as LLaVA [18, 21], Qwen-VL [1, 2], and InternVL [6, 7] effectively align visual features with LLMs. This paradigm has expanded from static images to video with models such as LLaVA-Video [50]

and InternVideo [36]. Despite progress in general visual question answering, these models still struggle with tasks requiring precise, localized visual reasoning [39, 40, 54].

Visual Limitations in MLLMs. MLLMs exhibit inherent limitations in fine-grained perception. They often fail on small-object recognition due to low-resolution visual encoders that lose critical details [40, 47]. Long video understanding remains challenging because restricted LLM context lengths prevent processing full sequences, forcing reliance on suboptimal sampling [54]. Additionally, most MLLMs lack strong temporal grounding capabilities, as they are not trained with explicit temporal signals or time-aware datasets, thus requiring specialized fine-tuning [4, 11]. Our work addresses all three limitations within a single framework.

Uncertainty in MLLMs. Uncertainty quantification is fundamental for building reliable models. In LLMs and MLLMs, entropy [29] is a key indicator of uncertainty, with higher entropy correlating with hallucinations and incorrect predictions [5, 14, 17]. Prior work has primarily used uncertainty for post-hoc hallucination detection, whereas we introduce a novel, training-free framework that employs entropy as a proactive signal to guide decision-making in complex visual tasks.

3 The Uncertainty-Guided (UG) Framework

In this section, we present the Uncertainty-Guided (UG) framework, a training-free approach that enhances MLLM performance on fine-grained perception tasks. We begin by establishing the core principle underlying our method (Sec. 3.1): actively minimizing a model's intrinsic uncertainty serves as an effective strategy for localizing salient visual input. After empirically validating this principle, we formalize the uncertainty metrics used to score visual inputs (Sec. 3.2) and detail how our simple "score-then-answer" mechanism applies to three distinct localization challenges (Sec. 3.3).

3.1 Principle: Minimizing Uncertainty to Localize Key Visual Information

Our framework rests on the hypothesis that an MLLM's intrinsic uncertainty can be minimized to localize the key visual information required to answer a query. Whereas prior work uses output entropy for post-hoc error detection, we use it as a proactive localization signal. To evaluate this idea, we conducted a motivating experiment on V^* Bench [40], which contains high-resolution images. Each image was partitioned into square crops with side length equal to one-sixth of $\min(\text{width}, \text{height})$. We fed each crop to LLaVA-OneVision-7B [18] with the original question, hypothesizing that crops containing salient information would yield *low-entropy (confident)* predictions, whereas irrelevant crops would yield *high-entropy (uncertain)* ones.

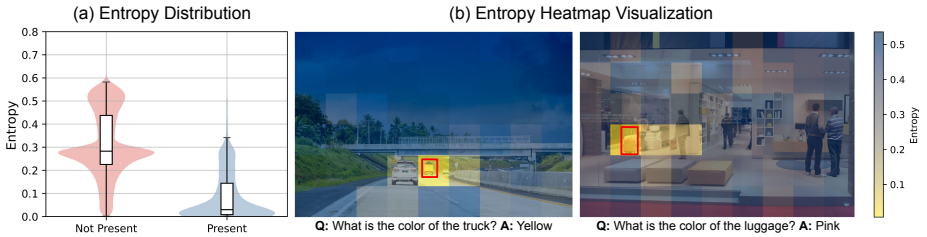

Fig. 1: (a) **Entropy distribution** comparing visual crops that include a target object (`Present`) versus those that do not (`Not Present`). (b) **Entropy heatmap** showing entropy variation across crops. Red box indicates the target object.

The results in Fig. 1 strongly validate this hypothesis. Fig. 1a shows a clear, statistically significant separation: `Present` crops containing the target object (e.g., the truck) exhibit substantially lower entropy than `Not Present` crops (e.g., background). Fig. 1b further visualizes this trend, where regions overlapping the target (red box) produce the lowest entropy. This strong correlation confirms that MLLM output uncertainty serves as a reliable proxy for visual relevance. Motivated by this observation, we reformulate fine-grained perception tasks as a search for the visual input that *minimizes model uncertainty*. Additional analysis and theoretical discussion can be found in the Appendix.

3.2 Uncertainty as a Scoring Function

A central component of our framework is quantifying the model's uncertainty for a visual input v and textual query q. We use two metrics derived from the output probability distribution p_i at each generation step i. For general-purpose relevance scoring, we employ *Average Entropy*. We compute Shannon entropy [29] for each generated token and average it across the sequence of length T:

$$\mathcal{H}(v, q) = -\frac{1}{T} \sum_{i=1}^{T} \sum_{j=1}^{N} p_{i,j} \log(p_{i,j}) \tag{1}$$

where N is the vocabulary size and $p_{i,j}$ is the probability of the jth vocabulary at step i. Lower entropy indicates lower uncertainty and thus higher alignment of the visual input with the query.

For tasks that require a binary decision (e.g., detecting whether an event is present), average entropy may be less informative since confident "yes" or "no" responses both yield low entropy. In such cases, we use a more targeted metric named *Binary Response Confidence (BRC)* score. It directly measures the model's certainty toward a positive confirmation by taking the difference between the probabilities of the "yes" and "no" tokens in the first generated distribution p_1:

$$BRC(v, q) = p_1(\text{“yes”}) - p_1(\text{“no”}) \tag{2}$$

This score provides a directional measure of uncertainty, where a high positive value indicates strong confidence in the event's presence.

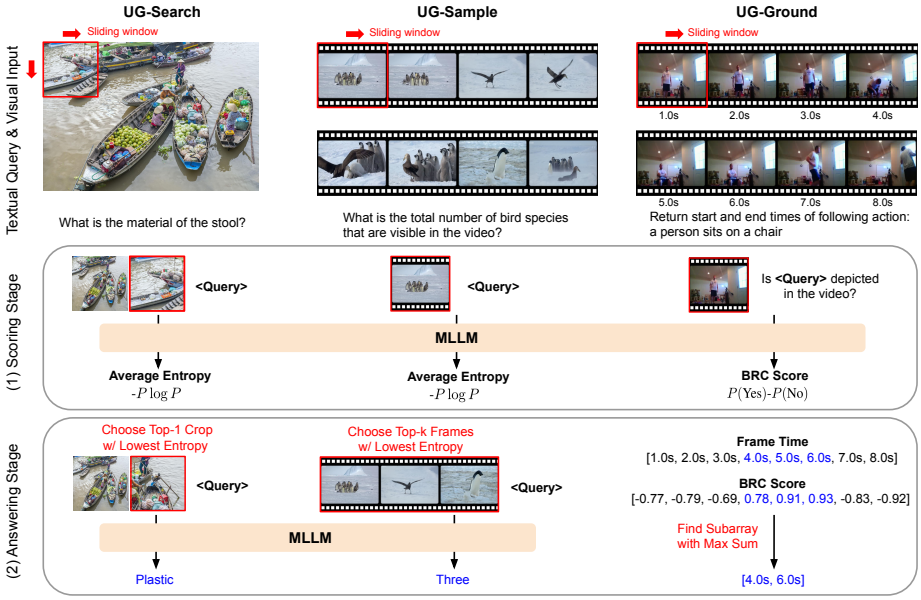

Fig. 2: Uncertainty-Guided (UG) Framework operates in two stages: (1) Scoring stage: candidate visual inputs (image crops or video frames) are scored using the MLLM's intrinsic uncertainty, measured by either Average Entropy or Binary Response Confidence (BRC) score. (2) Answering stage: the input with the lowest uncertainty is used for a final inference to generate the definitive answer.

3.3 Applications of the UG Framework

The UG framework's simple "score-then-answer" mechanism generalizes across localization tasks including Visual Search, Long Video Understanding, and Temporal Grounding, as shown in Fig. 2.

UG-Search: Visual Search in High-Resolution Images. To localize small objects in large images, UG-Search divides the image into a set of candidate crops using a sliding window. Each crop is scored by feeding it, along with the original image, to the MLLM and calculating its *Average Entropy* (Eq. (1)). This initial passes are solely for scoring. The single crop that yields the lowest entropy is selected as the most informative region, and the final answer is then generated based on this selected crop only. To improve efficiency in practice, we cache the original image's key-value (KV) states and reuse them across all scoring passes.

UG-Sample: Frame Sampling for Long Videos. We extend the same principle to long video sampling. To find the most relevant moments, UG-Sample treats each frame (or short window) as a candidate visual input. Each candidate is scored using its *Average Entropy*. Then, the top-k frames with the lowest entropy are selected, concatenated into a single context in temporal order, and used for final inference to answer the query.

UG-Ground: Temporal Grounding of Events. For temporal grounding, UG-Ground reframes the task as finding the most confident contiguous event segment. A sliding window is applied across the video, with each window scored using the *BRC score* (Eq. (2)) by querying whether the event occurs. This yields a sequence of confidence scores. Temporal grounding is thus reduced to finding the subarray with the maximum sum which can be solved efficiently in linear time using *Kadane's Algorithm* [3]. The resulting subarray's start and end indices correspond directly to the predicted timestamps.

4 Experiments

We conduct comprehensive experiments to evaluate the effectiveness, generality, and scalability of our Uncertainty-Guided (UG) framework. We assess performance on three challenging localization tasks: Visual Search (Sec. 4.1), Video Frame Sampling (Sec. 4.2), and Video Temporal Grounding (Sec. 4.3). Additional analyses and efficiency discussions are presented in Sec. 4.4 and Sec. 4.5. All evaluations are performed using *LMMs-Eval* [48] library for consistency and reproducibility. Implementation details are provided in the Appendix.

4.1 UG-Search on Visual Search

Table 1: Results of UG-Search on Visual Search and Standard QA. The value in subscript (Δ) indicates the performance gain from the baseline. We also report the average width ($\bar{W}$) and height ($\bar{H}$) of the images in each benchmark.

Model	Visual Search QA			Standard QA			
	V^* Bench	HR4K	HR8K	TextVQA	POPE	DocVQA	GQA
Image Size ($\bar{W} \times \bar{H}$)	*2246×1582*	*4023×3502*	*7431×5357*	*954×818*	*584×478*	*1776×2084*	*578×482*
LLaVA-OV-7B	74.4	64.9	58.4	73.8	88.4	87.1	62.3
w/ **UG-Search**$_\Delta$	86.9$_{12.5}$	70.1$_{5.2}$	68.9$_{10.5}$	75.1$_{1.3}$	89.7$_{1.3}$	88.2$_{0.9}$	63.0$_{0.7}$
Qwen2.5-VL-7B	64.9	60.1	53.1	76.1	86.3	93.1	60.4
w/ **UG-Search**$_\Delta$	81.7$_{16.8}$	74.9$_{14.8}$	66.3$_{13.2}$	79.7$_{3.6}$	86.7$_{0.4}$	94.2$_{0.9}$	60.7$_{0.3}$
InternVL2.5-8B	71.7	59.8	58.0	77.0	90.5	91.4	62.9
w/ **UG-Search**$_\Delta$	91.1$_{19.4}$	75.8$_{16.0}$	73.6$_{15.6}$	77.7$_{0.7}$	90.8$_{0.3}$	92.3$_{0.9}$	63.4$_{0.5}$

Setup. We process each image using a square sliding window. To account for image resolution, we set the crop size to one-sixth of the image for Visual Search QA and to one-half for Standard QA, with the stride set to half the crop size. These hyperparameters are fixed across all benchmarks and are not tuned for individual datasets. UG-Search is applied to several open-source MLLMs, including LLaVA-OneVision [18], Qwen2.5-VL [2], and InternVL-2.5 [6]. We evaluate on visual search benchmarks (V^* Bench [40] and HR-Bench [35]) and standard benchmarks (TextVQA [31], POPE [20], DocVQA [25], and GQA [13]). Benchmark details appear in the Appendix.

Results. Tab. 1 shows that UG-Search consistently improves performance across all base models and datasets. Gains are particularly strong for Visual Search

QA, which requires detecting small objects: InternVL2.5 improves by +19.4% on V^* Bench. This verifies that UG-Search effectively guides the model toward the correct region of interest. The method also generalizes to Standard QA tasks on regular-size images that still require fine-grained perception. For example, Qwen2.5-VL gains +3.6% on TextVQA and +0.9% on DocVQA, with LLaVA-OV and InternVL2.5 also showing consistent improvements. These results confirm that UG-Search enhances fine-grained perception across diverse visual QA settings without any training.

Tab. 2 compares UG-Search with specialized visual search methods built on the same MLLM backbones. Our approach outperforms ZoomEye [30], another training-free method, across all benchmarks. Methods such as TextCoT [23] and ViCrop [47], which use the MLLM to generate bounding boxes or rely on attention maps, provide limited gains on high-resolution datasets such as HR-Bench. Although Thyme [49] uses supervised fine-tuning and reinforcement learning to generate cropping code, UG-Search achieves superior results without training, underscoring the strength of our uncertainty-based approach.

Table 2: SOTA Comparison of UG-Search. The **bold** number represents the best performance.

Model	V^* Bench	HR4K	HR8K
LLaVA-OV-7B	74.4	64.9	58.4
w/ ZoomEye [30]	85.0	68.4	66.5
w/ **UG-Search**	**86.9**	**70.1**	**68.9**
Qwen2.5-VL-7B	64.9	60.1	53.1
w/ TextCoT [23]	67.0	60.6	50.6
w/ ViCrop [47]	69.6	63.3	51.9
w/ Thyme [49]	68.1	66.6	59.0
w/ **UG-Search**	**81.7**	**74.9**	**66.3**

Table 3: SOTA Comparison of UG-Sample. The **bold** number represents the best performance.

Model	V-MME	MLVU	LVB
LLaVA-OV-7B	53.9	58.6	54.3
w/ KFC [8]	55.4	**65.0**	55.6
w/ BOLT [22]	56.1	63.4	55.6
w/ AKS [32]	56.1	64.8	56.8
w/ FRAG [12]	56.3	64.9	57.3
w/ **UG-Sample**	58.6	62.0	**59.5**
w/ **UG-Sample**+AKS	**59.2**	**65.0**	59.4

4.2 UG-Sample on Video Frame Sampling

Setup. We next evaluate UG framework's ability to identify key query-relevant moments in videos. Following prior work [22,32], we select the top-8 frames from a pool of 256 uniformly sampled candidate frames, using a window size of one frame for entropy scoring. For baseline models, 8 frames are uniformly sampled over the entire video. Importantly, these hyperparameters are fixed across all benchmarks, demonstrating that the method does not rely on benchmark-specific optimization. In addition to previously utilized MLLMs (LLaVA-OneVision, Qwen2.5-VL, and InternVL-2.5), we also apply UG-Sample to video-specialized models (LLaVA-Video [50], InternVideo-2.5 [37]). We benchmark on Long Video QA (Video-MME [9], MLVU [53], and LongVideoBench [38]) and Short Video QA (EgoSchema [24], ActivityNet-QA [45], and NeXT-QA [41]).

Results. Tab. 4 shows that UG-Sample consistently improves over uniform sampling across all evaluated MLLMs. Notably, the method benefits not only long video tasks but also relatively short video benchmarks, illustrating the broad utility of uncertainty-guided frame selection. For example, Qwen2.5-VL gains

Table 4: Results of UG-Sample on Long Video and Short Video QA. The value in subscript (Δ) indicates the performance gain from the baseline. We also report the average video length of each benchmark.

Model	Long Video QA						Short Video QA		
	Video-MME (w/o sub.)				MLVU	LVB	EgoSch.	ANQA	NextQA
	Overall	Short	Medium	Long					
Video Length	*17min*	*1.3min*	*9min*	*41min*	*12min*	*8min*	*3min*	*2min*	*0.8min*
LLaVA-OV-7B	53.9	64.9	52.1	44.7	58.6	54.3	59.1	42.9	77.1
w/ **UG-Sample**$_\Delta$	58.6$_{4.7}$	69.7$_{4.8}$	58.4$_{6.3}$	47.7$_{3.0}$	62.0$_{3.4}$	59.5$_{5.2}$	60.3$_{1.2}$	44.1$_{1.2}$	77.5$_{0.4}$
Qwen2.5-VL-7B	53.7	62.7	51.7	46.8	53.9	52.7	53.7	44.0	69.6
w/ **UG-Sample**$_\Delta$	59.8$_{6.1}$	69.4$_{6.7}$	58.9$_{7.2}$	51.0$_{4.2}$	55.3$_{1.4}$	60.5$_{7.8}$	56.6$_{2.7}$	45.7$_{1.7}$	71.2$_{1.6}$
InternVL2.5-8B	57.8	68.1	56.4	48.8	61.6	52.8	49.2	42.6	76.6
w/ **UG-Sample**$_\Delta$	60.6$_{2.8}$	70.3$_{2.2}$	59.9$_{3.5}$	51.7$_{2.9}$	67.6$_{6.0}$	59.5$_{6.7}$	50.0$_{0.8}$	43.6$_{1.0}$	77.3$_{0.7}$
LLaVA-Video-7B	55.9	67.7	53.6	46.4	56.4	55.7	49.7	48.8	75.6
w/ **UG-Sample**$_\Delta$	59.7$_{3.8}$	70.6$_{2.9}$	60.1$_{6.5}$	48.3$_{1.9}$	61.9$_{5.5}$	58.1$_{2.4}$	51.2$_{1.5}$	49.0$_{0.2}$	77.5$_{2.9}$
InternVideo2.5-8B	54.2	62.9	52.7	47.0	59.4	51.2	56.4	48.0	76.5
w/ **UG-Sample**$_\Delta$	57.4$_{3.2}$	65.8$_{2.9}$	56.0$_{3.3}$	50.3$_{3.3}$	63.2$_{3.8}$	57.7$_{6.5}$	58.1$_{1.7}$	49.4$_{1.4}$	78.2$_{1.7}$

+6.1% on Video-MME and +7.8% on LongVideoBench. On short-video datasets, UG-Sample still provides meaningful improvements, boosting Qwen2.5-VL by +2.7% on EgoSchema and +1.6% on NextQA. As expected, gains are smaller for shorter videos, where uniform sampling is less likely to omit key moments. Overall, these results demonstrate that UG-Sample is a robust and effective strategy across diverse video lengths.

The comparison in Tab. 3 further demonstrates the strength of our approach. On Video-MME and LongVideoBench, UG-Sample alone outperforms all specialized baselines, including methods that rely on external visual-text similarity scores (BOLT [22], AKS [32]) and graph-based reasoning (KFC [8]). This suggests that an MLLM's intrinsic uncertainty constitutes a more semantically rich signal for frame relevance than external CLIP similarity or repeated binary prompting (FRAG [12]). On MLVU, however, UG-Sample underperforms relative to competing methods. We attribute this gap to MLVU's *Action Count* category, which demands whole-video coverage: because UG-Sample globally prioritizes the most salient frames, it may under-sample less-prominent target actions distributed across the video. To address this, we construct UG-Sample+AKS, a hybrid that replaces AKS's CLIP-based scoring with our entropy-based score while retaining AKS's segment-wise sampling strategy, thereby ensuring broad temporal coverage. This hybrid achieves state-of-the-art performance on Video-MME and MLVU, confirming that intrinsic model uncertainty and temporal coverage are complementary rather than competing objectives.

4.3 UG-Ground on Temporal Grounding

Setup. Finally, we evaluate our framework on video temporal grounding, which requires identifying the exact start and end times of an event within a video. For

Table 5: Results of UG-Ground on Video Temporal Grounding Benchmarks. The **bold** number represents the best performance. The value in subscript (Δ) indicates the improvement over the baseline. The **bold** number represents the best performance.

Model	Charades-STA (*0.5min*)				ActivityNet Captions (*3min*)			
	R@0.3	R@0.5	R@0.7	mIoU	R@0.3	R@0.5	R@0.7	mIoU
Instruction-tuned Methods								
TimeChat-7B	40.6	23.8	9.7	26.2	25.0	13.2	6.1	18.5
VTimeLLM-13B	55.3	34.3	14.7	34.6	44.8	29.5	14.2	31.4
TimeMarker-8B	73.5	51.9	26.9	48.4	-	-	-	-
Training-free Methods								
VTG-GPT	59.5	43.7	25.9	39.8	47.1	28.3	12.8	30.5
TFVTG	67.0	50.0	24.3	44.5	49.3	27.0	13.4	34.1
TAG	67.8	48.6	26.7	45.7	51.9	28.9	15.1	36.6
LLaVA-OV-7B	14.4	7.2	3.5	10.4	14.8	7.2	3.0	11.9
w/ **UG-Ground**$_\Delta$	69.0$_{54.6}$	45.7$_{38.5}$	25.9$_{22.4}$	46.8$_{36.4}$	53.4$_{38.6}$	29.2$_{22.0}$	15.5$_{12.5}$	37.2$_{25.3}$
Qwen2.5-VL-7B	61.1	43.7	22.9	41.1	19.0	9.5	4.1	14.3
w/ **UG-Ground**$_\Delta$	70.1$_{9.0}$	51.0$_{7.3}$	30.1$_{7.2}$	48.4$_{7.3}$	54.4$_{35.4}$	**31.4**$_{21.9}$	**16.5**$_{12.4}$	37.9$_{23.6}$
InternVL2.5-8B	11.7	4.1	1.9	9.5	6.0	2.4	1.0	5.7
w/ **UG-Ground**$_\Delta$	67.6$_{55.9}$	49.3$_{45.2}$	30.4$_{28.5}$	47.4$_{37.9}$	52.7$_{46.7}$	30.0$_{27.6}$	16.2$_{15.2}$	36.9$_{31.2}$
LLaVA-Video-7B	16.8	8.3	3.0	11.4	20.9	9.9	3.9	15.4
w/ **UG-Ground**$_\Delta$	67.5$_{50.7}$	47.2$_{38.9}$	29.2$_{26.4}$	47.2$_{35.8}$	**54.9**$_{34.0}$	30.8$_{20.9}$	**16.5**$_{12.6}$	38.3$_{22.9}$
InternVideo2.5-8B	50.2	32.0	14.4	32.3	18.9	9.5	4.1	14.0
w/ **UG-Ground**$_\Delta$	**74.7**$_{24.5}$	**52.5**$_{20.5}$	**32.6**$_{18.2}$	**51.0**$_{18.7}$	**54.9**$_{36.0}$	30.8$_{21.3}$	16.4$_{12.3}$	**38.5**$_{24.5}$

baseline, the input is capped at 64 frames due to the context-length constraints of MLLMs. For UG-Sample, the BRC score is computed using a 15-frame sliding window with stride 1, and this configuration is used for all datasets without benchmark-specific hyperparameter tuning. UG-Ground is applied to the same five MLLMs used in UG-Sample. Experiments are conducted on two standard benchmarks: Charades-STA [10] (around 0.5-minute videos) and ActivityNet Captions [15] (around 3-minute videos). Considering their average video durations, we sample videos at 3 FPS for Charades-STA and 1 FPS for ActivityNet Captions. Following prior works [4, 28], we report standard metrics including Recall@k at various IoU thresholds and mean Intersection over Union (mIoU).

Results. Tab. 5 shows that most MLLMs perform poorly at temporal grounding out of the box as they are not trained for such precise localization. However, UG-Ground dramatically enhances their capabilities, yielding large improvements across all models and datasets. For example, LLaVA-OV's mIoU on Charades-STA rises from 10.4 to 46.8 (+36.4), demonstrating the substantial impact of uncertainty-guided temporal localization.

Remarkably, this training-free method enables general-purpose MLLMs to outperform two distinct categories of temporal-specialized models: (1) instruction-tuned methods, such as VTimeLLM [11], TimeChat [28], and TimeMarker [4] that are explicitly fine-tuned on temporal reasoning data or with time-aware tokens; and (2) training-free methods, such as VTG-GPT [43], TFVTG [51], and TAG [16] that rely on an external pretrained VLM to generate image captions or compute frame-query similarity scores. These results demonstrate that lever-

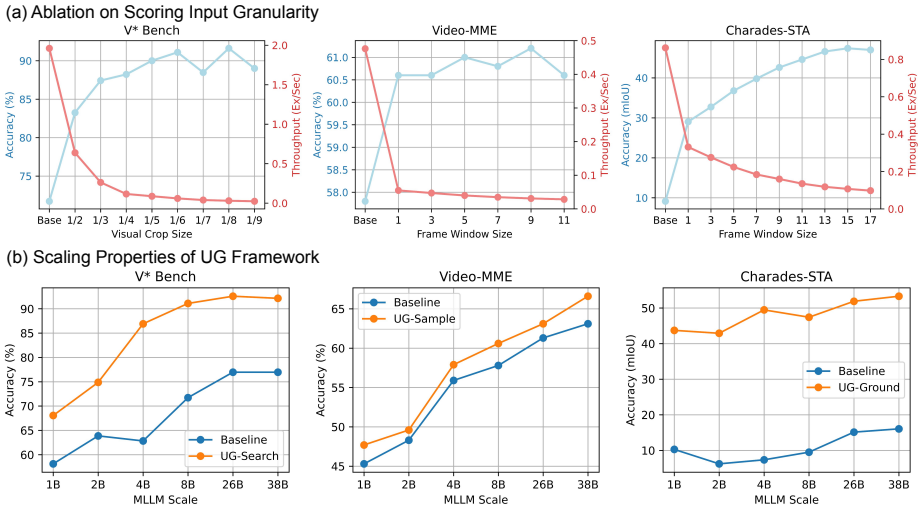

Fig. 3: (a) **Ablation on scoring input granularity** shows a clear trade-off between accuracy and throughput. `Base` denotes the baseline. (b) **Scaling properties of UG framework** show consistent performance gains as InternVL-2.5 model size increases.

aging the MLLM's own internal confidence signal via the BRC score constitutes a more principled and effective strategy for temporal localization than either fine-tuning on temporal reasoning data (VTimeLLM, TimeChat) or introducing time-aware tokens (TimeMarker). UG-Ground also consistently outperforms other training-free approaches such as TFVTG and TAG. Finally, applying UG-Ground to a video-specialized backbone like InternVideo2.5 achieves state-of-the-art performance, underscoring the strength and versatility of our unified grounding framework.

4.4 UG Framework Analysis

In this section, we conduct a series of analyses to better understand the behavioral properties of the UG framework. We first examine how scoring input granularity affects performance and efficiency, then evaluate scalability with respect to model size, followed by qualitative examples.

Ablation on Scoring Input Granularity. We first examine how the granularity of candidate visual inputs used during scoring (crop size or window size) affects accuracy and efficiency using InternVL2.5-8B. Throughput (examples/second) is measured on NVIDIA H100 GPUs. As shown in Fig. 3a, results exhibit a clear trade-off between accuracy and speed.

For UG-Search on V^* Bench, accuracy (light blue) peaks at finer granularity (crop size of 1/8), with gains saturating beyond 1/6. Throughput (red), however, declines sharply as granularity increases, dropping from roughly 0.6 at 1/2 to under 0.1 at 1/9. This shows that extremely fine crops yield diminishing accuracy gains while incurring substantial computational cost.

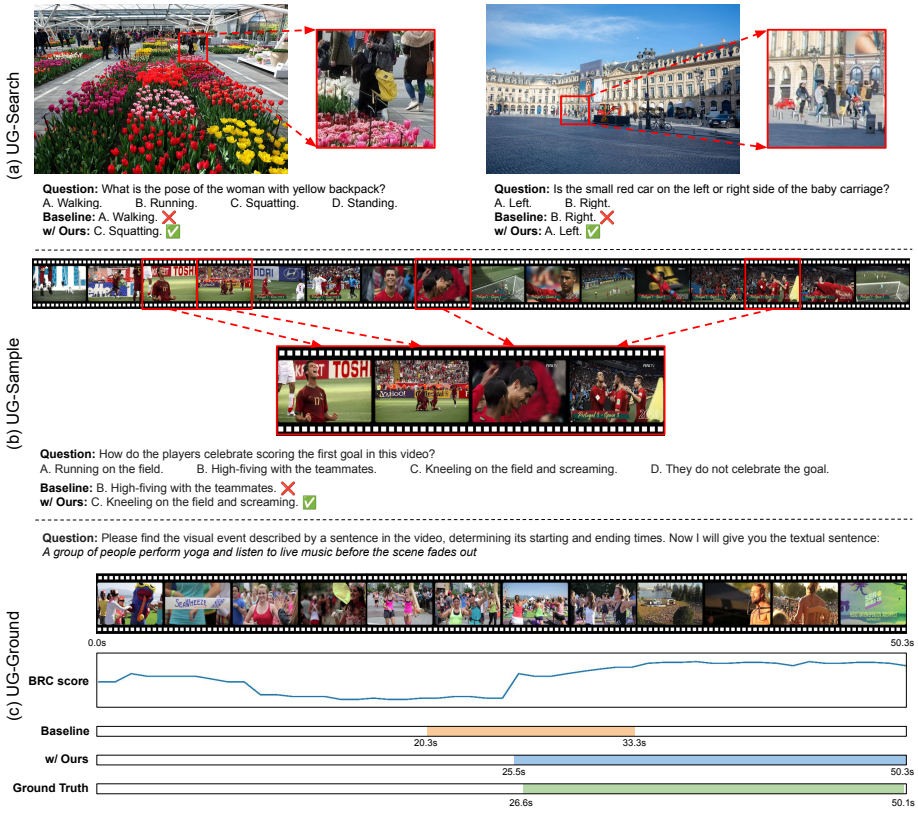

Fig. 4: Qualitative Results with InternVL2.5-8B as the baseline. (a) UG-Search localizes small target objects by selecting the lowest-entropy crop (red box). (b) UG-Sample identifies key semantic frames (red box) with the lowest entropy from a long video understanding. (c) UG-Ground pinpoints the correct event timeline by finding the peak in its BRC score sequence.

For video tasks, the optimal granularity varies by task. UG-Ground evaluated on Charades-STA benefits from larger temporal windows, achieving its best performance with a 15-frame window. In contrast, UG-Sample on Video-MME achieves near-optimal accuracy with a single frame, as capturing isolated key moments matters more than longer temporal context. In both cases, throughput decreases almost linearly with respect to window size.

Across all granularities, UG consistently surpasses the baseline, demonstrating robustness. In practice, users can select coarser inputs (e.g., 1/2 crops or 1-frame windows) for higher throughput, or finer inputs (e.g., 1/8 crops or 15-frame windows) for maximum accuracy, offering flexible control over the performance–efficiency trade-off.

Scaling Properties of UG Framework. Fig. 3b evaluates the scalability of the UG across the InternVL-2.5 model family, spanning 1B to 38B parame-

ters. Both the original baseline models and their UG-enhanced variants improve steadily as model size increases. Crucially, the benefits of UG remain stable and consistent across the entire scaling curve, and larger models achieve the highest absolute performance. This consistent improvement suggests that as MLLMs become larger and, presumably, better calibrated, their intrinsic uncertainty becomes an even more reliable signal that our framework can effectively harness.

Qualitative Results. Fig. 4 provides qualitative examples demonstrating how the UG framework isolates the most relevant visual information in noisy visual contexts, enabling correct predictions where baseline models fail. The examples are drawn from V^* Bench, Video-MME, and ActivityNet Captions, with additional qualitative results included in the Appendix.

In the examples for (a), UG-Search overcomes the baseline's tendency to be distracted by cluttered scenes. In the left image, it isolates the woman with the backpack to correctly identify her pose as "Squatting", while the baseline incorrectly guesses a common pose ("Walking"). Similarly, on the right image, it successfully localizes "small red car" and "baby carriage" to accurately resolve their spatial relationship.

The video example in (b) shows UG-Sample identifying the precise frames of a goal celebration. The baseline, likely processing irrelevant moments from a uniform sample, fails to understand the event correctly. Our method, however, filters the timeline to provide the necessary context relevant to the query, leading to the correct answer ("C. Kneeling on the field and screaming").

Finally, in (c), the baseline fails to identify the correct time segment for the "yoga and live music" event. Our UG-Ground method first transforms the video into a sequence of BRC scores, which exhibits a clear plateau of high confidence during the correct interval. This allows our method to accurately ground the event's duration, aligning almost perfectly with the ground truth.

4.5 Improving Efficiency of UG Framework

The main limitation of the UG framework is its increased runtime due to the multiple scoring passes required during evaluation. To address this, we introduce three practical strategies for improving efficiency: (1) decoupling the scoring and answering MLLMs, (2) leveraging external vision models as pre-filters, and (3) dynamically adjusting the stride for temporal windows. These approaches offer practical avenues for significantly reducing computational cost while maintaining strong performance. Moreover, the UG framework is inherently parallelizable: crops or frames can be processed independently, enabling substantial reductions in wall-clock time when batched or distributed across devices.

Decoupling the Scoring and Answering Models. We explore separating the MLLM used during the uncertainty-scoring stage from the model used for final answer generation. As shown in Tab. 6, all decoupled configurations yield clear gains over the no-scoring baseline (71.7% and 57.8% on V^* Bench and Video-MME, respectively), confirming that the scoring step itself is the primary driver of improvement. Across scoring model sizes, accuracy and throughput trade off monotonically: a 1B scorer achieves the highest throughput among UG

Table 6: Results with Decoupled Scoring and Answering Models. Using a larger MLLM for scoring improves accuracy, while a smaller scorer increases throughput during inference. Throughput is reported in example per second.

Answering MLLM	Scoring MLLM	UG-Search		UG-Sample	
		V^* Bench	Throughput	Video-MME	Throughput
InternVL2.5-8B	None	71.7	1.961	57.8	0.476
InternVL2.5-8B	InternVL2.5-1B	75.4	**0.139**	58.5	**0.107**
InternVL2.5-8B	InternVL2.5-2B	79.1	0.097	59.0	0.097
InternVL2.5-8B	InternVL2.5-4B	89.0	0.078	60.1	0.070
InternVL2.5-8B	InternVL2.5-8B	91.1	0.060	60.6	0.054
InternVL2.5-8B	InternVL2.5-26B	**92.0**	0.037	**61.8**	0.036

variants (0.139 on V^* Bench) but the lowest accuracy (75.4%), while a 26B scorer delivers the best accuracy on both benchmarks (V^* Bench: 92.0%, Video-MME: 61.8%) at the cost of reduced throughput (0.037). Notably, the 4B scorer reaches 89.0% on V^* Bench within 2.1 points of self-scoring with the 8B model while maintaining a 1.3× throughput advantage, representing a favorable accuracy–efficiency operating point. The consistent accuracy gain from larger scorers aligns with the finding in Sec. 4.4 that larger, better-calibrated models produce more reliable entropy signals. This decoupling strategy opens practical directions such as using a lightweight distilled scorer or a cascaded scoring scheme to match large-scorer accuracy at lower cost.

Table 7: Results of Utilizing External Vision Models as Pre-filters.

(a) UG-Search with Pre-filters. UG-Search is applied to smaller subsets of visual candidates proposed by YOLOv8 and SAM2.

Method	V^*	Throughput
LLaVA-OV-7B	74.4	1.951
w/ **UG-Search**	**81.2**	0.420
+ YOLOv8 Pre-filter	74.9	**0.678**
+ SAM2 Pre-filter	78.2	0.228

(b) UG-Sample with Pre-filters. Top-32 frames are pre-filtered by SigLIP and DINOv2 before UG-Sample.

Method	V-MME	Throughput
LLaVA-OV-7B	53.9	1.048
w/ **UG-Sample**	**58.6**	0.130
+ DINOv2 Pre-filter	56.4	**0.232**
+ SigLIP Pre-filter	57.3	0.210

Using External Pre-filters to Improve Efficiency. We incorporate lightweight, off-the-shelf vision models as efficient *pre-filters* to reduce the number of candidates that the UG framework must score. These models identify a smaller and more promising subset of regions or frames, thereby lowering computational cost.

For UG-Search, we use region proposals from YOLOv8 [33] and segmentation masks from SAM2 [27] to construct a reduced set of candidate crops. As shown in Tab. 7a, full UG-Search achieves the highest accuracy (81.2%), albeit with a significant throughput reduction relative to the vanilla MLLM (0.420 vs. 1.951). Applying YOLOv8 as a pre-filter improves throughput among UG-Search variants (0.678) but yields only marginal gains over the vanilla MLLM (74.9% vs. 74.4%). SAM2 improves accuracy to 78.2% but incurs the highest latency (0.228),

making its efficiency–accuracy trade-off less favorable than full UG-Search. We attribute this to the query-agnostic nature of generic detectors: without access to the language query, they tend to miss small or context-dependent targets, a finding also reported in ViCrop [47]. Overall, these results suggest that in complex scenes, exhaustive UG-Search remains more reliable than generic region-selection pre-filters.

In contrast, pre-filtering offers a meaningful throughput–accuracy trade-off for video tasks. As shown in Tab. 7b, both DINOv2 [26] and SigLIP [46] pre-filters substantially improve throughput over full UG-Sample (0.232 and 0.210 vs. 0.130) while retaining most of the accuracy gain over the baseline. DINOv2 achieves the highest throughput (1.8× over UG-Sample) but at a larger accuracy cost (56.4% vs. 58.6%). SigLIP offers a better accuracy–efficiency balance: it narrows the accuracy gap to only 1.3 points below full UG-Sample (57.3% vs. 58.6%) while still delivering a 1.6× throughput improvement, and remains well above the no-sampling baseline (53.9%). This indicates that query-aware semantic pre-filtering can recover a significant portion of the compute overhead introduced by UG-Sample with only modest accuracy degradation.

Table 8: Results of Video Frame and Stride Adjustment.

(a) UG-Sample evaluated on Video-MME. We ablate Frame $\in \{1, 9\}$ and Stride $\in \{1, 3, 5, 7, 9\}$ during scoring stage.

Method	V-MME	Throughput
InternVL2.5-8B	57.8	0.476
w/ Frame=1, Stride=1	60.6	0.054
w/ Frame=9, Stride=1	61.1	0.025
w/ Frame=9, Stride=3	**62.2**	0.051
w/ Frame=9, Stride=5	61.9	0.067
w/ Frame=9, Stride=7	61.3	0.086
w/ Frame=9, Stride=9	60.6	**0.135**

(b) UG-Ground evaluated on Charades-STA. We ablate Frame $\in \{7, 15\}$ and Stride $\in \{1, 3, 5, 7, 9, 11\}$ during scoring stage.

Method	Charades	Throughput
InternVideo2.5-8B	32.3	1.370
w/ Frame=7, Stride=1	45.5	0.205
w/ Frame=15, Stride=1	**51.0**	0.150
w/ Frame=15, Stride=3	**51.0**	0.403
w/ Frame=15, Stride=5	50.9	0.599
w/ Frame=15, Stride=7	50.6	0.787
w/ Frame=15, Stride=9	50.2	0.877
w/ Frame=15, Stride=11	49.7	**1.064**

Dynamic Stride Adjustment for Frame-Window. We introduce dynamic stride control to balance temporal resolution with computational efficiency in both UG-Sample and UG-Ground. Increasing the stride reduces window overlap and thus lowers the number of forward passes required during the scoring stage of video-based UG methods.

Tab. 8a reports results on Video-MME using InternVL2.5-8B. The default UG-Sample configuration corresponds to Frame=1 and Stride=1. We then fix the window size to 9 frames and vary the stride, selecting the center frame of each window (e.g., the 5th frame in a 9-frame window) for Top-K frame selection. Increasing the stride from 1 to 9 increases throughput dramatically (0.025 vs. 0.135) while still outperforming the baseline (60.6% vs. 57.8%), showing that UG-Sample is highly robust to temporal sparsification.

A similar pattern holds for UG-Ground. The stride controls the temporal density of uncertainty scoring, governing the trade-off between localization pre-

cision and computational cost. As shown in Tab. 8b, the default configuration (Frame=15, Stride=1) delivers the strongest performance at 51.0 mIoU. Increasing the stride introduces only minor degradation: a stride of 3 reaches the same peak accuracy, and even a stride of 11 maintains strong performance at 49.7 mIoU. Throughput, however, improves substantially (0.150 $\rightarrow$ 1.064), adding only modest overhead relative to the baseline while preserving a large accuracy advantage (49.7 vs. 32.3 mIoU).

5 Conclusion and Limitation

In this work, we introduced the Uncertainty-Guided (UG) framework, a training-free approach that leverages an MLLM's intrinsic uncertainty to address challenging fine-grained perception tasks, including Visual Search, Long Video Understanding, and Temporal Grounding. Our experiments show that this simple yet principled strategy enables standard MLLMs to achieve performance competitive with heavily fine-tuned, task-specific systems, demonstrating that uncertainty minimization is an effective and efficient mechanism for guiding multimodal reasoning.

Limitation. The primary limitation of our framework is the increased inference cost introduced by the scoring stage, which requires multiple forward passes. However, this overhead can be alleviated using the efficiency strategies discussed in Sec. 4.5. More broadly, the UG framework aligns with the growing paradigm of test-time scaling, trading additional computation for enhanced capability, but differs in that its computation is inherently parallelizable: image crops or video frames are independent units that can be batched or distributed across devices to reduce wall-clock latency. Although our current implementation performs scoring sequentially due to resource and stability constraints, a fully parallelized design would substantially accelerate inference. We leave this direction, along with further efficiency optimizations, for future work.

Acknowledgements.

This work was partially funded by the ERC (853489 - DEXIM) and the Alfried Krupp von Bohlen und Halbach Foundation, which we thank for their generous support. We also gratefully acknowledge support from the German Research Foundation (DFG) via SFB 1233 "Robust Vision: Inference Principles and Neural Mechanisms," TP A2 (Project No. 276693517), and from Hi! PARIS and the ANR/France 2030 program (ANR-23-IACL-0005).

References

1. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A frontier large vision-language model with versatile abilities. arXiv preprint arXiv:2308.12966 (2023)

2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
3. Bentley, J.: Programming pearls: algorithm design techniques. Communications of the ACM **27**(9), 865–873 (1984)
4. Chen, S., Lan, X., Yuan, Y., Jie, Z., Ma, L.: Timemarker: A versatile video-llm for long and short video understanding with superior temporal localization ability. arXiv preprint arXiv:2411.18211 (2024)
5. Chen, S., Xiong, M., Liu, J., Wu, Z., Xiao, T., Gao, S., He, J.: In-context sharpness as alerts: An inner representation perspective for hallucination mitigation. In: ICLR (2024)
6. Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271 (2024)
7. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: CVPR (2024)
8. Fang, B., Wu, W., Wu, Q., Song, Y., Chan, A.B.: Threading keyframe with narratives: Mllms as strong long video comprehenders. arXiv preprint arXiv:2505.24158 (2025)
9. Fu, C., Dai, Y., Luo, Y., Li, L., Ren, S., Zhang, R., Wang, Z., Zhou, C., Shen, Y., Zhang, M., et al.: Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In: CVPR (2025)
10. Gao, J., Sun, C., Yang, Z., Nevatia, R.: Tall: Temporal activity localization via language query. In: ICCV (2017)
11. Huang, B., Wang, X., Chen, H., Song, Z., Zhu, W.: Vtimellm: Empower llm to grasp video moments. In: CVPR (2024)
12. Huang, D.A., Radhakrishnan, S., Yu, Z., Kautz, J.: Frag: Frame selection augmented generation for long video and long document understanding. arXiv preprint arXiv:2504.17447 (2025)
13. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: CVPR (2019)
14. Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., et al.: Language models (mostly) know what they know. arXiv preprint arXiv:2207.05221 (2022)
15. Krishna, R., Hata, K., Ren, F., Fei-Fei, L., Carlos Niebles, J.: Dense-captioning events in videos. In: ICCV (2017)
16. Lee, J.S., Lee, S., Ahn, J., Choi, Y., Lee, J.H.: Tag: A simple yet effective temporal-aware approach for zero-shot video temporal grounding. arXiv preprint arXiv:2508.07925 (2025)
17. Leng, S., Zhang, H., Chen, G., Li, X., Lu, S., Miao, C., Bing, L.: Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In: CVPR (2024)
18. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024)
19. Li, G., Xu, J., Zhao, Y., Peng, Y.: Dyfo: A training-free dynamic focus visual search for enhancing lmms in fine-grained visual understanding. In: CVPR (2025)
20. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., Wen, J.R.: Evaluating object hallucination in large vision-language models. In: EMNLP (2023)

21. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: CVPR (2024)
22. Liu, S., , Zhao, C., Xu, T., Ghanem, B.: Bolt: Boost large vision-language model without training for long-form video understanding. In: CVPR (2025)
23. Luan, B., Feng, H., Chen, H., Wang, Y., Zhou, W., Li, H.: Textcot: Zoom in for enhanced multimodal text-rich image understanding. arXiv preprint arXiv:2404.09797 (2024)
24. Mangalam, K., Akshulakov, R., Malik, J.: Egoschema: A diagnostic benchmark for very long-form video language understanding. In: NeurIPS (2023)
25. Mathew, M., Karatzas, D., Jawahar, C.: Docvqa: A dataset for vqa on document images. In: WACV (2021)
26. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
27. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. In: ICLR (2025)
28. Ren, S., Yao, L., Li, S., Sun, X., Hou, L.: Timechat: A time-sensitive multimodal large language model for long video understanding. In: CVPR (2024)
29. Shannon, C.E.: A mathematical theory of communication. The Bell system technical journal **27**(3), 379–423 (1948)
30. Shen, H., Zhao, K., Zhao, T., Xu, R., Zhang, Z., Zhu, M., Yin, J.: Zoomeye: Enhancing multimodal llms with human-like zooming capabilities through tree-based image exploration. In: EMNLP (2025)
31. Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., Rohrbach, M.: Towards vqa models that can read. In: CVPR (2019)
32. Tang, X., Qiu, J., Xie, L., Tian, Y., Jiao, J., Ye, Q.: Adaptive keyframe sampling for long video understanding. In: CVPR (2025)
33. Varghese, R., Sambath, M.: Yolov8: A novel object detection algorithm with enhanced performance and robustness. In: 2024 International conference on advances in data engineering and intelligent computing systems (ADICS) (2024)
34. Wang, J., Gao, Y., Sang, J.: Valid: Mitigating the hallucination of large vision language models by visual layer fusion contrastive decoding. arXiv preprint arXiv:2411.15839 (2024)
35. Wang, W., Ding, L., Zeng, M., Zhou, X., Shen, L., Luo, Y., Yu, W., Tao, D.: Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In: AAAI (2025)
36. Wang, Y., Li, K., Li, Y., He, Y., Huang, B., Zhao, Z., Zhang, H., Xu, J., Liu, Y., Wang, Z., et al.: Internvideo: General video foundation models via generative and discriminative learning. arXiv preprint arXiv:2212.03191 (2022)
37. Wang, Y., Li, X., Yan, Z., He, Y., Yu, J., Zeng, X., Wang, C., Ma, C., Huang, H., Gao, J., et al.: Internvideo2.5: Empowering video mllms with long and rich context modeling. arXiv preprint arXiv:2501.12386 (2025)
38. Wu, H., Li, D., Chen, B., Li, J.: Longvideobench: A benchmark for long-context interleaved video-language understanding. In: NeurIPS (2024)
39. Wu, J., Liu, W., Liu, Y., Liu, M., Nie, L., Lin, Z., Chen, C.W.: A survey on video temporal grounding with multimodal large language model. arXiv preprint arXiv:2508.10922 (2025)
40. Wu, P., Xie, S.: V?: Guided visual search as a core mechanism in multimodal llms. In: CVPR (2024)

41. Xiao, J., Shang, X., Yao, A., Chua, T.S.: Next-qa: Next phase of question-answering to explaining temporal actions. In: CVPR (2021)
42. Xiao, Y., Wang, W.Y.: On hallucination and predictive uncertainty in conditional language generation. In: ACL (2021)
43. Xu, Y., Sun, Y., Xie, Z., Zhai, B., Du, S.: Vtg-gpt: Tuning-free zero-shot video temporal grounding with gpt. Applied Sciences **14**(5), 1894 (2024)
44. Yu, S., Jin, C., Wang, H., Chen, Z., Jin, S., Zuo, Z., Xu, X., Sun, Z., Zhang, B., Wu, J., et al.: Frame-voyager: Learning to query frames for video large language models. In: ICLR (2024)
45. Yu, Z., Xu, D., Yu, J., Yu, T., Zhao, Z., Zhuang, Y., Tao, D.: Activitynet-qa: A dataset for understanding complex web videos via question answering. In: AAAI (2019)
46. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: ICCV (2023)
47. Zhang, J., Khayatkhoei, M., Chhikara, P., Ilievski, F.: Mllms know where to look: Training-free perception of small visual details with multimodal llms. ICLR (2025)
48. Zhang, K., Li, B., Zhang, P., Pu, F., Cahyono, J.A., Hu, K., Liu, S., Zhang, Y., Yang, J., Li, C., et al.: Lmms-eval: Reality check on the evaluation of large multimodal models. arXiv preprint arXiv:2407.12772 (2024)
49. Zhang, Y.F., Lu, X., Yin, S., Fu, C., Chen, W., Hu, X., Wen, B., Jiang, K., Liu, C., Zhang, T., et al.: Thyme: Think beyond images. arXiv preprint arXiv:2508.11630 (2025)
50. Zhang, Y., Wu, J., Li, W., Li, B., Ma, Z., Liu, Z., Li, C.: Video instruction tuning with synthetic data. arXiv preprint arXiv:2410.02713 (2024)
51. Zheng, M., Cai, X., Chen, Q., Peng, Y., Liu, Y.: Training-free video temporal grounding using large-scale pre-trained models. In: ECCV (2024)
52. Zheng, Z., Yang, M., Hong, J., Zhao, C., Xu, G., Yang, L., Shen, C., Yu, X.: Deepeyes: Incentivizing" thinking with images" via reinforcement learning. arXiv preprint arXiv:2505.14362 (2025)
53. Zhou, J., Shu, Y., Zhao, B., Wu, B., Liang, Z., Xiao, S., Qin, M., Yang, X., Xiong, Y., Zhang, B., et al.: Mlvu: Benchmarking multi-task long video understanding. In: CVPR (2025)
54. Zou, H., Luo, T., Xie, G., Lv, F., Wang, G., Chen, J., Wang, Z., Zhang, H., Zhang, H., et al.: From seconds to hours: Reviewing multimodal large language models on comprehensive long video understanding. arXiv preprint arXiv:2409.18938 (2024)