OMNICOMPASS: A HOLISTIC BENCHMARK FOR OMNI-MODAL VIDEO GROUNDING AND UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

With the rapid development of multimodal large language models (MLLMs), evaluating their video understanding capabilities has become increasingly important. Existing video understanding benchmarks cover only part of the evaluation dimensions, limiting systematic diagnosis of omni-modal video understanding. We introduce **OmniCompass**, a holistic benchmark for omni-modal video grounding and understanding. OmniCompass defines **two evaluation tracks**: **OmniCompass-VL** assesses vision-language models on tasks requiring visual evidence, while **OmniCompass-AV** assesses omni-modal models on tasks requiring joint audio-visual evidence. Each track includes **three complementary tasks**: temporal grounding (TG), spatial grounding (SG), and question answering (QA), assessing *when* a referenced target appears, *where* it is, and *what* happens. QA covers five perception and three reasoning categories in OmniCompass-VL, as well as three perception and two reasoning categories in OmniCompass-AV. OmniCompass-VL contains 820 videos and 2,460 instances across various domains and **three duration types**: short, medium, and long. A narrative- and audio-rich subset forms OmniCompass-AV with 240 videos and 720 instances. Each instance is evaluated under **two paired settings**: text-only and text+image, specifying the same target through language alone and visual-language guidance. All instances are **manually annotated** with multi-round quality verification. Evaluations of **50** MLLMs show that open-source models still lag behind proprietary counterparts overall, particularly in audio-visual understanding. Precise temporal and spatial grounding remains challenging even for leading models, and their predictions still exhibit inconsistencies across the paired target specifications.

1 INTRODUCTION

Human perception uses linguistic or visual cues to identify targets, attend to visual and auditory evidence across space and time, and form goal-directed judgments. Multimodal large language models (MLLMs) jointly process images, videos, audio, and text, supporting multimodal understanding and reasoning (Bai et al., 2025a; Xu et al., 2025b; Google, 2025; Kimi Team et al., 2026). Recent MLLMs can revisit relevant temporal clips (Yang et al., 2026; Zeng et al., 2026), inspect task-relevant spatial evidence (Wang et al., 2026; Meng et al., 2026), and selectively acquire audio-visual evidence during reasoning (Xing et al., 2026; Chen et al., 2026). These advances motivate evaluating omni-modal video grounding and understanding. This requires identifying targets from textual or visual cues, grounding them spatiotemporally, and integrating audio-visual evidence to understand events.

With the rapid development of MLLMs, how to comprehensively evaluate their performance in complex video understanding tasks has become an important problem. Table 1 compares existing benchmarks across four dimensions. **(i) Target Specification** defines how targets are presented: through only textual queries (Fu et al., 2025) or with reference images (Ma et al., 2026). **(ii) Video Evidence** distinguishes tasks resolved from video frames (Chen et al., 2025a) from those requiring synchronized audio-visual content (Hong et al., 2026). **(iii) Task Coverage** specifies the output: temporal grounding (TG) and spatial grounding (SG) predict target-specific intervals and regions (Lei et al., 2021; Cheng et al., 2026), while question answering (QA) predicts answers (Fu et al., 2026). **(iv) Video Duration** determines the temporal scope: short-video benchmarks emphasize fine-grained motion and local dynamics (Li et al., 2024b), whereas long-video benchmarks test long-range dependencies and information integration across temporally distant events (Zhou et al., 2025a).

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.

Table 1: **Comparison of existing video understanding benchmarks.** Target Specification indicates whether the target is specified using text alone (*i.e.*, Text) or text with an image (*i.e.*, Text+Image); Video Evidence indicates whether correct outputs depend on visual or native audio content. TG/SG/QA denote temporal grounding, spatial grounding, and question answering. Short/Med./Long denote video length of < 90s, $[90, 300)$s, and ≥ 300s. #Vid./#Inst. are video/instance counts; M/A/H denote manual/automatic/hybrid annotation; Len. is the average video length in seconds.

Benchmarks	Target Specification		Video Evidence		Task Coverage			Video Duration			#Vid.	#Inst.	Anno.	Len.(s)
	Text	Text+Image	Visual	Audio	TG	SG	QA	Short	Med.	Long				
Video Benchmarks														
Charades-STA (Gao et al., 2017)	✓	✗	✓	✗	✓	✗	✗	✓	✗	✗	1,334	3,720	M	30.0
TVQA (Lei et al., 2018)	✓	✗	✓	✗	✗	✗	✓	✓	✗	✗	2,179	15,253	M	11.2
TVQA+ (Lei et al., 2020)	✓	✗	✓	✗	✓	✓	✓	✓	✗	✗	403	2,821	M	61.5
VidSTG (Zhang et al., 2020)	✓	✗	✓	✗	✓	✓	✗	✓	✗	✗	732	10,303	M	28.0
NExT-QA (Xiao et al., 2021)	✓	✗	✓	✗	✗	✗	✓	✓	✗	✗	1,000	8,564	A	44.0
HC-STVG (Tang et al., 2021)	✓	✗	✓	✗	✓	✓	✗	✓	✗	✗	1,160	1,160	M	20.0
QVHighlights (Lei et al., 2021)	✓	✗	✓	✗	✓	✗	✗	✗	✓	✗	1,529	1,542	M	149.6
EgoSchema (Mangalam et al., 2023)	✓	✗	✓	✗	✗	✗	✓	✗	✓	✗	5,063	5,063	H	180.0
TempCompass (Liu et al., 2024b)	✓	✗	✓	✗	✗	✗	✓	✓	✗	✗	410	7,540	H	11.4
MVBench (Li et al., 2024b)	✓	✗	✓	✗	✗	✗	✓	✓	✗	✗	3,641	4,000	A	16.0
MLVU (Zhou et al., 2025a)	✓	✗	✓	✗	✗	✗	✓	✗	✓	✓	1,334	2,593	H	720.0
Video-MME (Fu et al., 2025)	✓	✗	✓	✗	✗	✗	✓	✓	✓	✓	900	2,700	M	1017.9
LongVideoBench (Wu et al., 2024)	✓	✗	✓	✗	✗	✗	✓	✓	✓	✓	3,763	6,678	M	473.0
CG-Bench (Chen et al., 2025a)	✓	✗	✓	✗	✗	✗	✓	✗	✗	✓	1,219	12,129	M	1624.4
V-STaR (Cheng et al., 2026)	✓	✗	✓	✗	✓	✓	✓	✓	✓	✓	1,000	2,094	H	110.2
IV-Bench (Ma et al., 2026)	✗	✓	✓	✗	✗	✗	✓	✗	✓	✓	967	2,585	M	537.0
OmniCompass-VL (Ours)	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	820	2,460	M	354.0
Audio-Visual Benchmarks														
AVUT (Yang et al., 2025)	✓	✗	✓	✓	✗	✗	✓	✓	✗	✗	2,662	11,609	H	67.8
Daily-Omni (Zhou et al., 2025b)	✓	✗	✓	✓	✗	✗	✓	✓	✗	✗	684	1,197	H	42.8
OmniVideoBench (Li et al., 2026)	✓	✗	✓	✓	✗	✗	✓	✓	✓	✓	628	1,000	H	384.2
WorldSense (Hong et al., 2026)	✓	✗	✓	✓	✗	✗	✓	✓	✓	✓	1,662	3,172	M	141.1
JointAVBench (Chao et al., 2026)	✓	✗	✓	✓	✗	✗	✓	✓	✓	✓	1,046	2,853	H	97.2
Video-MME-v2 (Fu et al., 2026)	✓	✗	✓	✓	✗	✗	✓	✓	✓	✓	800	3,200	M	624.0
OmniCompass-AV (Ours)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	240	720	M	427.5

Despite this progress, existing benchmarks have three limitations in systematically diagnosing omni-modal video grounding and understanding. **(i) Limited video-evidence coverage.** Benchmarks typically assess either visual information (Chen et al., 2025a) or synchronized audio-visual content (Hong et al., 2026). In news broadcasts, some queries concern visible scenes; others require aligning spoken reports with people or events. Both settings are needed to assess visual perception and audio-visual integration. **(ii) Inflexible target-specification settings.** Most benchmarks use only text (Fu et al., 2026), while IV-Bench combines reference images with text (Ma et al., 2026). Practical targets may be specified verbally or with an image, yet these benchmarks do not evaluate the same target under both settings. **(iii) Fragmented task coverage.** Most benchmarks focus on QA (Li et al., 2024b) or grounding (Gao et al., 2026); audio-visual benchmarks rarely cover the latter (Li et al., 2026; Zhou et al., 2025b). In basketball games, users may ask **when** a referenced player scores, **where** the player is on the court, and **what** happens. Joint evaluation distinguishes failures in temporal retrieval, spatial localization, and semantic understanding.

To address these gaps, we introduce **OmniCompass**, a holistic benchmark for omni-modal video grounding and understanding. As shown in Figure 1, OmniCompass has four key features: **(i) Complementary Video Evidence.** OmniCompass-VL evaluates understanding from visual content, while OmniCompass-AV additionally requires audio. **(ii) Flexible Target Specification.** Each instance pairs *Text* and *Text+Image* settings, specifying the same target through language alone or with a reference image. The video, task semantics, and ground truth are the same. We introduce Cross-Specification Consistency Scores (CS) to jointly measure task accuracy and prediction consistency across these settings. **(iii) Unified Task Coverage.** Both tracks provide TG, SG, and QA tasks per video. The VL track covers five perception and three reasoning categories, whereas the AV track covers three perception and two reasoning categories. **(iv) Diverse Durations and Content.** OmniCompass-VL contains 820 videos and 2,460 instances across short, medium, and long durations. It combines public sports, traffic, surveillance, and documentary videos with 240 self-collected movie, TV, variety, and news videos. This subset forms OmniCompass-AV with 720 instances.

We evaluate **50 MLLMs**, including proprietary and open-source models. The best proprietary model exceeds the best open-source model by **3.2%** in $\mathrm{CS}^{\mathrm{Avg}}$ on VL track and **32.8%** on AV track. Strong QA performance coexists with weak grounding and cross-reference inconsistency, even when

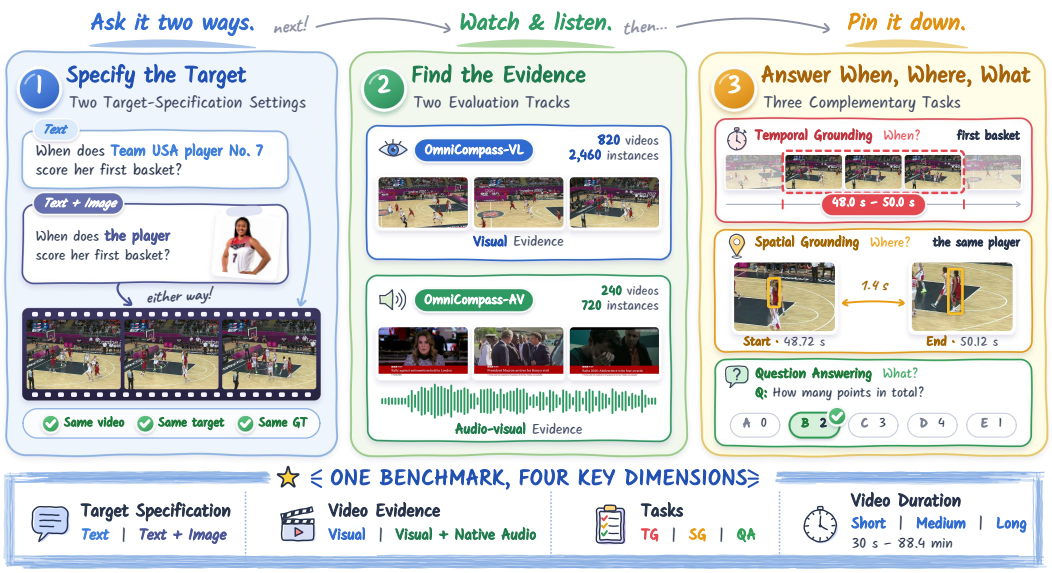


Figure 1: **Overview of OmniCompass.** The benchmark spans four dimensions: target specification, video evidence, task coverage, and video duration. Paired *Text*/*Text+Image* specifications are evaluated on the same targets. OmniCompass-VL contains 820 videos and 2,460 instances, while the 240-video OmniCompass-AV subset contains 720 instances with native audio. Both tracks cover temporal grounding, spatial grounding, and question answering.

aggregate accuracies are similar across settings. Visual target specifications do not consistently help: *Text+Image* improves the single-setting average for only 11 of 36 models. Further analyses show that timestamps substantially improve model performance across tasks and tracks, while audio consistently improves QA and TG performance on AV track. Future MLLMs therefore need stronger cross-modal alignment and more reliable audio-aware grounding. Our contributions are as follows:

1. **A Holistic Evaluation Framework.** OmniCompass unifies target-guided omni-modal video grounding and understanding across four dimensions: paired *Text* and *Text+Image* specifications, VL/AV evidence tracks, TG/SG/QA tasks, and short-to-long videos. We introduce CS to jointly assess task correctness and consistency across paired target specifications.
2. **A Diverse, Rigorously Verified Benchmark.** OmniCompass spans 820 videos, 80.6 hours, seven domains, and 52 subcategories, with 240 videos shared between VL and AV tracks. Construction required about 1,800 human-hours for annotator training, calibration, annotation, expert adjudication, and revision. All instances undergo multi-round quality verification.
3. **A Diagnostic Evaluation of MLLMs.** We evaluate 50 different MLLMs, revealing an overall advantage for proprietary models, especially in audio-visual understanding. Strong QA performance coexists with weak grounding and cross-specification inconsistency. Further analyses examine the effects of target specifications, video duration, timestamps, and input modalities.

2 Related Work

Multimodal Large Language Models (MLLMs) for Video Understanding. MLLMs connect language models with multimodal encoders for perception and reasoning (Liu et al., 2024a; Chen et al., 2024; Bai et al., 2025b; Xu et al., 2025a; Google, 2025; Li et al., 2025; An et al., 2026). LLaVA-OneVision (Li et al., 2024a) and LLaVA-Video (Zhang et al., 2024) unify image-video modeling, while VideoLLaMA3 (Zhang et al., 2025) and InternVL3.5 (Wang et al., 2025b) strengthen video training and reasoning. Qwen3-VL (Bai et al., 2025a) improves long-context video understanding through temporal alignment, while Qwen3-Omni (Xu et al., 2025b) jointly processes vision, audio, and language. Agentic extensions also explore tool-assisted video reasoning to selectively acquire task-relevant visual and audio evidence (Yang et al., 2026; Meng et al., 2026; Xing et al., 2026). In brief, modern omni-modal video understanding identifies referenced targets and grounds evidence in time and space by integrating visual and audio information.

Video Understanding Benchmarks. Table 1 compares existing benchmarks across four design dimensions. **(i) Target Specification** defines how the target is presented. Most benchmarks use textual questions or instructions (Fu et al., 2025; Cheng et al., 2026), while IV-Bench (Ma et al., 2026) pairs text with a reference image for visual specification. **(ii) Video Evidence** distinguishes visual and audio-visual requirements. Vision-language benchmarks assess appearance, motion, and temporal relations. TempCompass (Liu et al., 2024b) evaluates temporal perception, while MotionBench (Hong et al., 2025) targets fine-grained motion. WorldSense (Hong et al., 2026) and JointAVBench (Chao et al., 2026) require synchronized visual and native-audio evidence. **(iii) Task Coverage** defines the expected output. Question-answering benchmarks (Fu et al., 2025; Li et al., 2026) produce semantic responses, while grounding benchmarks require explicit evidence localization. OmniGround evaluates temporal intervals and spatial regions (Gao et al., 2026), and V-STaR (Cheng et al., 2026) jointly assesses TG, SG, and QA. **(iv) Video Duration** sets the temporal scope for evidence integration. MVBench emphasizes local actions and temporal dynamics (Li et al., 2024b), whereas LongVideoBench (Wu et al., 2024), MLVU (Zhou et al., 2025a), and LVBench (Wang et al., 2025a) stress long-context management across distant moments. Existing benchmarks cover selected dimensions or fixed combinations, limiting diagnosis of target identification, grounding, and omni-modal reasoning across different target and evidence settings.

OmniCompass unifies these dimensions through paired target-specification settings, complementary evidence tracks, and TG, SG, and QA across video durations. It supports diagnosis of target identification, spatio-temporal grounding, and event understanding.

3 The OmniCompass Benchmark

OmniCompass evaluates MLLMs in omni-modal video grounding and understanding across the dimensions in Table 1. Figure 1 illustrates their integration in the benchmark workflow. Each instance uses paired *Text* and *Text+Image* specifications of the same target. **OmniCompass-VL** evaluates vision-language models on visually resolvable tasks, while **OmniCompass-AV** evaluates omni-modal models using joint visual and native-audio evidence. Both tracks cover TG, SG, and QA across short ($30\text{s} \leq t < 90\text{s}$), medium ($90\text{s} \leq t < 300\text{s}$), and long ($t \geq 300\text{s}$) videos.

3.1 Evaluation Formulation and Metrics

Two Target-Specification Settings. OmniCompass uses paired $\widehat{35}$extitText and *Text+Image* settings to test whether models consistently localize and understand targets across reference modalities. Targets include entities, actions, attributes, and scenes. In *Text*, a textual description $\mathbf{r}^{\text{text}}$ alone specifies the target, such as *"When does the USA player wearing No. 33 complete the dunk?"*. In *Text+Image*, a textual expression $\mathbf{r}^{\text{text}'}$ and a reference image $\mathbf{r}^{\text{image}}$ jointly specify the target. For example, *"When does the player complete the dunk?"* requires the image to resolve which player is referenced. Thus, $\mathbf{r} = \mathbf{r}^{\text{text}}$ or $\mathbf{r} = (\mathbf{r}^{\text{text}'}, \mathbf{r}^{\text{image}})$. Each pair shares the same video, target, task semantics, and ground truth, allowing direct comparison across target specifications.

Two Evaluation Tracks. The two tracks share the same target-specification and task contracts but differ in the video evidence available to the model. Let $\mathcal{F}_K(\mathcal{V})$ denote K uniformly sampled video frames paired with their timestamps, and let $\mathcal{A}(\mathcal{V})$ denote the synchronized audio track. Given a target specification $\mathbf{r} \in \{\mathbf{r}^{\text{text}}, \mathbf{r}^{\text{image}}\}$, **OmniCompass-VL** receives $\mathbf{x}_{\text{VL}} = (\mathcal{F}_K(\mathcal{V}), \mathbf{r})$, whereas **OmniCompass-AV** receives $\mathbf{x}_{\text{AV}} = (\mathcal{F}_K(\mathcal{V}), \mathcal{A}(\mathcal{V}), \mathbf{r})$. VL queries are answerable from visual evidence, while AV queries require information from the synchronized audio.

Cross-Specification Consistency Scores (CS). Paired *Text* and *Text+Image* settings allow us to assess **whether a model correctly identifies the same target across referential modalities**. We introduce **CS** to jointly measure task correctness and consistency between paired predictions, so agreement alone does not guarantee a high score. For each task $q \in \{\text{TG}, \text{SG}, \text{QA}\}$, let N_q denote the number of evaluated pairs, with superscripts T and TI denoting the *Text* and *Text+Image* settings. For TG, let g_i^{T} and g_i^{TI} be the tIoUs of the predicted intervals with the ground-truth interval, and a_i the tIoU between the two predictions. We define $\text{CS}_{\text{TG}} = \frac{1}{N_{\text{TG}}} \sum_{i=1}^{N_{\text{TG}}} \min\{g_i^{\text{T}}, g_i^{\text{TI}}, a_i\}$. For SG, let $g_{i,k}^{\text{T}}$ and $g_{i,k}^{\text{TI}}$ be the IoUs between predicted and ground-truth boxes at the start and end boundaries $k \in \{s, e\}$. Let $a_{i,k}$ denote the IoU between the paired predictions at boundary k. We

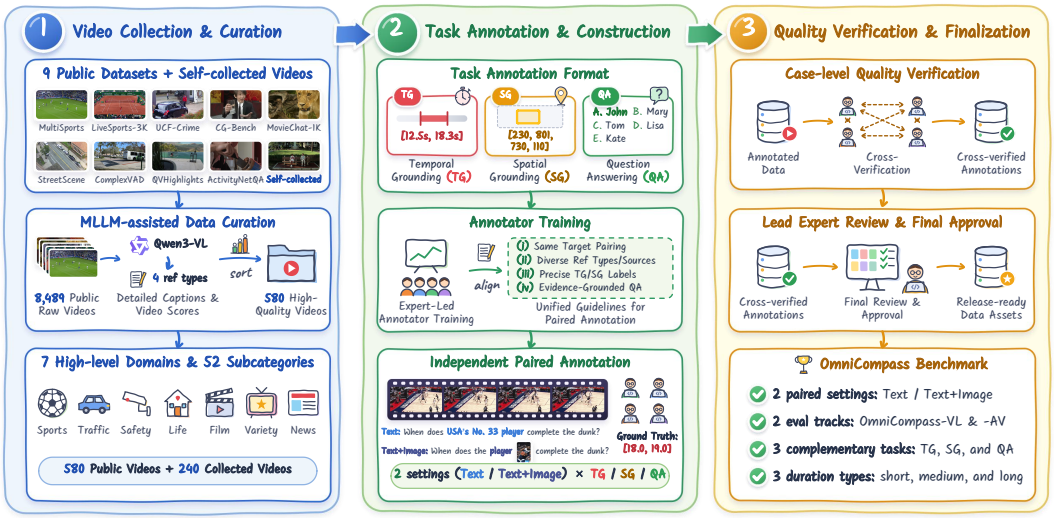


Figure 2: **Data construction pipeline.** OmniCompass requires about **1,800 human-hours** for rigorous annotation, verification, and revision. **Stage 1:** We screen 8,489 videos from nine public datasets, retain 6,946 after duration and referential-cue filtering, and select 580 to complement 240 self-collected videos. **Stage 2:** Four trained annotators independently construct paired *Text* and *Text+Image* TG, SG, and QA instances. AV cases include native audio. **Stage 3:** Annotators cross-verify and revise cases before final acceptance and correction by the lead expert, yielding 820 videos across seven domains, paired settings, two tracks, three tasks, and three duration groups.

define $\mathrm{CS}_{\mathrm{SG}} = \frac{1}{2N_{\mathrm{SG}}} \sum_{i=1}^{N_{\mathrm{SG}}} \sum_{k\in\{s,e\}} \min\{g_{i,k}^{\mathrm{T}}, g_{i,k}^{\mathrm{TI}}, a_{i,k}\}$. For QA, with predicted answers $\hat{y}_i^{\mathrm{T}}$ and $\hat{y}_i^{\mathrm{TI}}$ and ground-truth answer y_i, we define $\mathrm{CS}_{\mathrm{QA}} = \frac{1}{N_{\mathrm{QA}}} \sum_{i=1}^{N_{\mathrm{QA}}} \mathbf{1}\{\hat{y}_i^{\mathrm{T}} = \hat{y}_i^{\mathrm{TI}} = y_i\}$.

3.2 Benchmark Construction

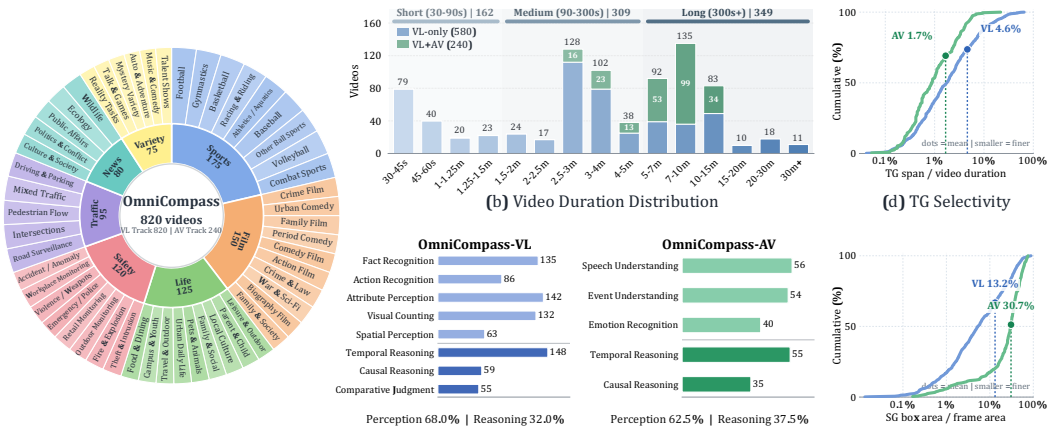
Constructing OmniCompass requires approximately **1,800 human-hours** for paired annotation, targeted revision, and multi-round quality verification, following the pipeline in Figure 2.

Stage 1: Video Collection and Curation. We construct a pool of **8,489 candidate videos** from **nine public datasets** and supplement it with **240 self-collected videos** from films, TV series, variety shows, and news programs. The resulting pool covers seven high-level domains. Appendix Table C1 details the selection scope, characteristics, and usage terms of the public sources.

We use `Qwen3-VL-8B-Instruct` to profile each public candidate with structured `entity`, `action`, `attribute`, and `scene` cues. Excluding videos shorter than 30 seconds or missing any of the four referential cue types leaves **6,946 eligible videos**. We rank these using four predefined scoring criteria (Table C2 in Appendix), then shortlist them under domain, duration, source, and subcategory constraints. Manual review retains **580 high-quality videos**, which we combine with the 240 self-collected videos to obtain 820 videos for OmniCompass-VL. The same 240 videos form OmniCompass-AV. Appendix C details profiling, filtering, and scoring.

Stage 2: Task Annotation and Construction. A lead expert provides two days of task-specific training to four professional annotators before full-scale annotation. The team then follows four principles to ensure consistent target specification and task evidence across TG, SG, and QA.

(i) Same-Target Pairing. Each *Text* and *Text+Image* pair shares the same video, target, task semantics, and ground truth. In *Text+Image*, the textual expression requires the reference image to identify the target. **(ii) Diverse Reference Types and Sources.** References cover `entity`, `action`, `attribute`, and `scene` cues using video crops or semantically matched external images. Each image identifies the target without revealing the annotated event or answer evidence. **(iii) Precise TG/SG Labels.** TG intervals tightly cover complete target-related events, while SG boxes localize the same target at both interval boundaries. **(iv) Evidence-Grounded QA Design.** Following CG-Bench (Chen et al., 2025a), annotators mark a supporting clue interval and construct five mutually exclusive choices with one evidence-supported answer and four plausible distractors.



(a) Video Domains: 7 Domains & 52 Subcategories (b) Video Duration Distribution (c) QA Types: Perception and Reasoning Tasks (d) TG Selectivity (e) SG Granularity

Figure 3: **Dataset statistics of OmniCompass.** **(a)** 820 videos span 7 domains and 52 subcategories, each with at least ten videos. **(b)** Fifteen duration bins span short (30–90 s), medium (90–300 s), and long ($\geq$ 300 s) ranges. Stacked colors distinguish 580 VL-only and 240 shared VL–AV videos. **(c)** Bars report counts for eight OmniCompass-VL and five OmniCompass-AV QA types. Lighter and darker shades distinguish Perception and Reasoning. **(d)** Cumulative TG-span ratios show temporal selectivity. **(e)** Cumulative SG-area ratios show spatial granularity. Dots mark track means.

The annotators first conduct a **pilot annotation** on 30 VL and 30 AV videos spanning all seven domains and three duration types. Each video receives one TG, one SG, and one QA instance under paired *Text* and *Text+Image* settings, yielding **180 task instances**. The pilot exposes inconsistencies in target pairing, temporal boundaries, spatial target consistency, and QA evidence. The lead expert reviews these cases with the team, resolves ambiguities, and refines the guidelines. The annotators then independently construct the remaining instances under the calibrated protocol. Guideline training, pilot calibration, and initial task construction account for approximately **960 annotator-hours**.

Stage 3: Quality Verification and Finalization. All annotated cases undergo case-level quality verification before release. Four annotators independently cross-check annotations from their peers against the Stage 2 principles: same-target pairing, reference diversity and non-leakage, precise TG/SG labels, and evidence-grounded QA design. For AV cases, they also verify native-audio dependence and audio-visual alignment. They correct clear annotation errors, recheck the revised cases, and submit the confirmed cases to the lead expert. The expert resolves remaining inconsistencies, makes necessary corrections, performs final acceptance, and audits the release-ready data. These steps require approximately **320 annotator-hours** and **500 lead-expert-hours**. Appendix D details the process; Tables D1–D2 report verification, and Table D3 reports effort.

3.3 Benchmark Statistics

Video Diversity and Temporal Scale. OmniCompass contains **820 videos** across **7 domains and 52 subcategories**, with 10–43 videos per subcategory and a mean of 15.8. The videos total **80.6 hours**, ranging from 30 seconds to 88.4 minutes, with a mean duration of 354.0 seconds. Figure 3(a,b) shows 162 short, 309 medium, and 349 long videos across 15 duration bins, including 580 VL-only videos and 240 videos shared by both tracks for matched-video evaluation.

Evaluation Breadth and Annotation Granularity. Unified TG, SG, and QA annotations yield **2,460 VL and 720 AV task instances**. Each video contributes one instance to each task, giving 820 instances per task in VL and 240 in AV. Figure 3(c) shows eight VL and five AV QA types. Perception accounts for 68.0% and 62.5%, while reasoning accounts for 32.0% and 37.5%, respectively. The 820 videos span seven domains and 52 subcategories, with 10–43 videos per subcategory. QA counts range from 55 to 148 across VL types and from 35 to 56 across AV types, providing broad coverage without concentrating on a single question category. Figure 3(d,e) summarizes grounding granularity: TG intervals average **4.6%** of VL and **1.7%** of AV video duration, while SG boxes average **13.2%** and **30.7%** of frame area, respectively. Paired *Text* and *Text+Image* settings test target identification, while the shared AV subset supports matched-video comparisons. Appendix Figure C4 reports reference types and image sources across domains.

Table 2: **Results on the OmniCompass-VL.** Models are evaluated on TG, SG, and QA without audio using uniformly sampled, timestamped frames. Performance is reported as *Text*/*Text+Image*, while **CS** requires the paired predictions to be both correct and mutually consistent. **Average** averages TG m_{tIoU}, SG m_{sIoU}, and QA Accuracy per specification, while CS^{Avg} averages the three CS scores. The best and second-best results within each group are shown in **bold** and underlined, respectively.

Models	Frames	Temporal Grounding			Spatial Grounding			Question Answering		Overall	
		R.@0.5	m_{tIoU}	CS^{TG}	AP@0.5	m_{sIoU}	CS^{SG}	Accuracy	CS^{QA}	Average	CS^{Avg}
Proprietary MLLMs											
Gemini-3.8-Flash	64	**33.8/32.7**	**32.6/33.5**	**27.1**	37.5/**41.6**	37.7/**41.8**	32.3	73.3/**76.6**	**69.1**	**47.9/50.6**	**42.8**
Gemini-3.1-Pro	64	24.9/24.9	25.6/26.6	19.3	34.9/35.2	36.5/37.3	31.8	71.4/73.6	65.7	44.5/45.9	38.9
Seed-2.1-Pro-260915	64	27.9/29.8	29.3/29.9	23.7	**40.9**/38.9	**40.6**/39.6	**34.9**	73.4/74.3	67.4	47.8/47.9	42.0
Seed-2.1-Turbo-260628	64	29.9/29.4	30.5/30.2	25.0	35.4/35.4	36.3/36.3	30.0	74.1/74.1	67.7	47.0/46.9	40.9
Seed-2.0-Pro-260215	64	30.7/31.0	30.6/30.9	24.9	30.6/31.7	34.9/35.3	29.4	**74.5**/73.0	67.3	46.6/46.4	40.5
Qwen3.8-Max	64	29.3/30.3	30.7/32.2	26.7	30.1/32.1	33.9/34.0	26.7	72.4/72.5	67.2	45.7/46.2	40.2
Qwen3.8-Flash	64	28.5/30.1	29.5/30.6	24.1	22.6/24.4	28.8/29.3	22.6	69.2/71.0	62.4	42.5/43.6	36.4
GPT-6-Astra	49	30.0/30.3	30.9/31.7	26.3	34.6/35.9	37.8/38.3	32.7	69.4/65.6	59.4	46.0/45.2	39.5
GPT-5.6-Sol	49	25.2/28.7	27.2/30.5	22.2	31.7/29.8	34.8/33.8	28.0	69.5/69.1	62.4	43.8/44.5	37.5
GPT-5.6-Terra	49	20.1/23.7	24.1/26.3	19.3	19.4/20.9	26.3/26.6	20.1	68.4/69.5	62.0	39.6/40.8	33.8
GPT-5.6-Luna	49	21.0/21.3	24.7/24.8	19.2	27.3/28.9	32.2/33.1	26.8	65.5/64.6	56.7	40.8/40.8	34.2
GPT-5.5	49	23.3/24.5	26.0/27.5	21.3	20.6/20.5	26.4/26.3	20.7	67.9/70.2	62.7	40.1/41.3	34.9
Open-Source MLLMs											
Kimi-K3	64	23.5/24.6	27.1/27.7	22.6	**30.0/31.6**	33.3/**35.2**	**28.4**	**73.0/72.9**	**67.6**	**44.5/45.3**	**39.6**
Qwen3.5-397B-A17B	64	**28.9/28.5**	28.6/28.7	23.1	28.2/29.6	**33.4**/33.7	28.0	67.7/69.6	60.6	43.3/44.0	37.2
Qwen3.5-122B-A10B	64	27.2/27.2	**29.7/29.1**	**24.3**	20.0/19.4	25.2/25.2	20.8	62.6/62.5	55.4	39.2/38.9	33.5
Qwen3-VL-235B-A22B-Instruct	64	27.0/25.4	28.5/26.0	21.6	18.6/18.8	26.0/25.0	20.9	68.4/65.3	61.0	41.0/38.8	34.5
GLM-5.3-Flash	64	24.4/23.0	25.5/24.4	18.3	15.7/14.6	22.9/23.5	14.6	68.3/68.7	59.4	38.9/38.9	30.8
DeepSeek-V4.1-Flash	64	15.7/17.4	19.1/19.4	12.4	11.5/11.6	18.1/17.9	10.5	66.2/64.6	56.6	34.5/34.0	26.5
MiMo-V2.5	64	17.8/18.6	21.6/20.8	13.8	14.2/15.3	21.9/22.3	13.6	62.0/60.8	49.0	35.2/34.7	25.5
Qwen3.8-27B	32	17.5/19.1	20.3/22.2	16.0	16.2/19.0	23.9/25.4	17.2	64.8/62.6	55.7	36.3/36.7	29.6
Qwen3.6-35B-A3B	32	18.7/18.8	19.8/19.7	14.4	20.9/19.1	26.8/26.4	20.9	59.5/59.9	50.2	35.4/35.3	28.5
Qwen3.6-27B	32	20.5/18.8	21.5/20.4	15.7	14.4/15.8	20.7/22.2	15.3	62.5/61.2	54.8	34.9/34.6	28.6
Qwen3.5-35B-A3B	32	17.3/16.4	19.5/18.7	13.7	13.3/13.4	21.2/21.9	15.8	61.3/60.9	51.9	34.0/33.8	27.1
Qwen3.5-27B	32	19.4/18.0	22.0/21.1	17.4	11.3/11.5	18.6/19.1	15.3	61.5/58.4	53.9	34.0/32.9	28.9
Qwen3.5-9B	32	17.4/14.4	20.2/16.6	13.3	16.8/16.8	22.3/22.3	17.7	58.5/55.2	50.1	33.7/31.4	27.0
Qwen3.5-4B	32	16.6/12.4	19.5/15.3	12.6	13.0/11.6	18.9/17.1	13.8	55.7/54.0	46.6	31.4/28.8	24.3
Qwen3-VL-30B-A3B-Instruct	32	16.8/12.4	19.5/14.9	11.7	15.2/12.9	22.1/19.6	15.6	54.1/49.6	43.4	31.9/28.0	23.6
Qwen3-VL-8B-Instruct	32	13.2/11.1	16.3/13.9	11.3	13.8/12.4	20.4/19.3	15.9	59.9/56.7	51.2	32.2/29.9	26.1
Qwen3-VL-4B-Instruct	32	13.8/10.5	16.8/13.0	10.3	6.6/5.6	13.4/12.2	9.3	57.3/56.0	50.2	29.2/27.1	23.3
InternVL3.5-38B	32	13.0/10.5	16.3/13.6	10.1	0.2/0.1	2.5/2.5	1.4	61.3/58.4	52.4	26.7/24.8	21.3
Keye-VL-2.0-30B-A3B	32	13.4/11.5	15.8/13.7	11.2	2.7/3.3	11.5/12.2	8.2	60.2/61.8	53.0	29.2/29.2	24.1
TimeLens2-8B	32	20.7/18.2	20.8/18.5	15.3	2.4/2.6	9.1/8.0	6.3	59.5/57.0	51.0	29.8/27.8	24.2
TimeLens-8B	32	22.4/18.5	21.4/18.1	15.5	2.4/2.8	8.2/8.7	5.9	57.7/55.5	49.5	29.1/27.4	23.6
LLaVA-OneVision-2-8B	32	15.4/10.7	15.8/11.2	8.4	0.9/1.1	7.1/6.7	4.6	57.9/52.4	48.0	27.0/23.4	20.4
LLaVA-Video-7B	32	1.6/2.1	4.3/3.1	1.5	0.0/0.0	0.5/0.5	0.1	55.0/50.7	45.6	19.9/18.1	15.7
LLaVA-OneVision-7B	32	2.6/1.7	4.6/2.4	1.4	0.0/0.0	2.0/1.6	0.8	50.7/47.9	42.3	19.1/17.3	14.8

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTING

Evaluated Models. We evaluate **50 models** across the VL and AV tracks. These include proprietary Gemini (Google, 2025), Seed (Seed, 2026), and GPT series (OpenAI, 2026), alongside open-source models such as Kimi-K3 (Kimi Team et al., 2026) and Qwen series (Team, 2026) across sizes. We use official APIs with default settings for proprietary models and models with at least 100B parameters, while evaluating smaller models on NVIDIA H100 GPUs.

Implementation Details. (i) Frame sampling: We uniformly sample 64 frames for proprietary models and open-source models with over 100B total parameters, and 32 for smaller models. GPT-series models use 49 frames to accommodate the *Text+Image* reference within the 50-image limit. VITA-1.5-7B follows its official 16-frame limit. Following TimeLens (Zhang et al., 2026), each frame is immediately preceded by its timestamp in seconds. **(ii) Input modalities:** OmniCompass-VL excludes audio, while OmniCompass-AV adds the native audio following WorldSense (Hong et al., 2026) and Video-MME-v2 (Fu et al., 2026). In *Text+Image*, the reference image follows these inputs and immediately precedes the task prompt. Appendix E provides details for experimental settings.

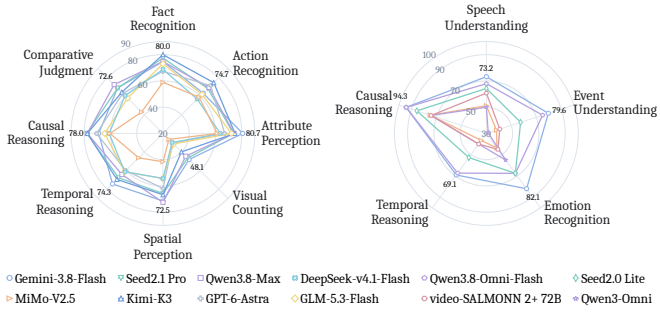
4.2 MAIN RESULTS

Tables 2 and 3 report TG, SG, and QA evaluations under *Text* and *Text+Image*. Figure 4 provides category-wise QA comparisons across selected models in both tracks. Appendix A provides additional results across video duration, domain, and target granularity.

Table 3: **Results on the OmniCompass-AV.** Models are evaluated on TG, SG, and QA using uniformly sampled, timestamped frames with synchronized audio.

Models	Frames	Temporal Grounding			Spatial Grounding			Question Answering		Overall	
		R.@0.5	m_{tIoU}	CS^{TG}	AP@0.5	m_{sIoU}	CS^{SG}	Accuracy	CS^{QA}	Average	CS^{Avg}
Proprietary Omni-modal MLLMs											
Gemini-3.8-Flash	64	**46.9/48.1**	**46.4/45.9**	**40.6**	**62.8/63.2**	**59.8/59.3**	**53.5**	**80.3/80.8**	**78.2**	**62.2/62.0**	**57.4**
Gemini-3.1-Pro	64	15.1/17.6	16.6/17.7	9.0	54.8/56.1	52.0/52.1	45.8	72.4/71.1	64.9	47.0/47.0	39.9
Qwen3.8-Omni-Flash	64	26.3/28.5	29.2/30.9	21.8	60.1/59.5	56.8/56.0	50.9	75.9/78.9	73.2	54.0/55.3	48.6
Qwen3.5-Omni-Plus	64	18.0/14.9	21.0/21.1	14.0	61.8/60.8	55.5/55.6	50.5	71.1/73.2	68.4	49.2/50.0	44.3
Qwen3.5-Omni-Flash	64	6.7/7.9	10.6/10.1	5.7	49.2/51.3	44.2/47.0	34.7	63.6/64.0	58.0	39.5/40.4	32.8
Seed-2.0-Lite-260428	64	13.3/8.3	16.9/13.2	10.6	56.3/52.9	53.3/51.8	45.5	71.3/69.2	63.8	47.1/44.7	39.9
Seed-2.0-Mini-260428	64	4.2/3.3	6.4/3.8	2.1	27.1/25.4	28.6/28.8	16.5	63.8/60.0	53.3	32.9/30.8	24.0
Open-Source Omni-modal MLLMs											
MiMo-V2.5	64	**6.3/8.6**	**9.4/8.7**	3.1	**29.2/33.2**	**34.9/36.0**	**23.4**	**59.2/58.7**	47.3	**34.5/34.5**	**24.6**
Qwen3-Omni-30B-A3B	32	3.8/3.3	5.5/5.0	**3.5**	24.2/24.6	27.4/29.8	17.3	53.3/55.0	47.9	28.7/30.0	22.9
Qwen2.5-Omni-7B	32	1.7/0.0	1.3/0.6	0.2	2.1/12.1	11.7/22.4	7.0	47.5/50.0	43.8	20.1/24.3	17.0
Qwen2.5-Omni-3B	32	1.3/2.1	1.4/2.5	0.8	0.4/2.5	1.9/7.6	0.7	48.8/50.0	45.0	17.4/20.0	15.5
MiniCPM-o-4.5-8B	32	0.0/0.0	1.9/1.4	0.4	1.3/1.7	10.4/9.0	5.0	52.1/52.9	47.1	21.5/21.1	17.5
video-SALMONN 2+ 72B	32	3.8/1.3	5.4/2.3	1.6	0.0/0.0	1.9/2.3	1.0	56.3/54.2	**50.8**	21.2/19.6	17.8
video-SALMONN 2+ 7B	32	2.9/2.1	3.1/1.8	1.0	0.0/0.0	0.9/0.1	0.1	47.1/51.3	42.9	17.0/17.7	14.7
Ola-7B	32	1.7/0.8	2.2/1.3	0.0	1.7/3.3	6.4/10.3	2.7	42.1/42.5	37.5	16.9/18.0	13.4
VideoLLaMA2.1-7B-AV	32	1.3/0.0	1.9/0.7	0.4	3.3/4.6	15.1/16.1	8.6	27.1/29.2	23.3	14.7/15.3	10.8
VITA-1.5-7B	16	0.0/0.8	0.5/1.2	0.2	0.0/0.0	1.1/1.3	0.5	34.2/34.6	27.5	11.9/12.4	9.4

OmniCompass-VL. Table 2 reports Gemini-3.8-Flash as the overall leader, achieving **42.8%** in CS^{Avg}, while Kimi-K3 is the top-performing open-source model at **39.6%**. **(i) Grounding and Fine-Grained Perception Lag Behind Reasoning.** 29 of 36 models exceed 50% in CS^{QA}, peaking at 69.1%. The best CS^{TG} and CS^{SG} are only 27.1% and 34.9%. Figure 4(a) shows that Visual Counting ranks last in CS^{QA} for all eight plotted models, at 26.4%–48.1%. Spatial Perception averages 61.0%, below 69.8% for Causal Reasoning. **(ii) Visual References Benefit Proprietary Models and Grounding.** *Text+Image* improves average performance for 11 of the 36 models: 8 of 12 proprietary models and 3 of 24 open-source models. Among the open-source models with fewer than 100B parameters, only Qwen3.8-27B improves. It gains 1.9% and 1.5% in TG m_{tIoU} and SG m_{sIoU} but loses 2.2% in QA accuracy. Visual cues may aid target matching while omitting useful textual detail. **(iii) Larger Models Do Not Lead Every Task.** Within Qwen3.5, 397B exceeds 122B in CS^{Avg}, 37.2% versus 33.5%, but trails in CS^{TG}, 23.1% versus 24.3%. Likewise, 27B trails 9B in CS^{SG}, 15.3% versus 17.7%. **(iv) Cross-Reference Consistency Remains Limited.** Across 36 models, CS^{TG}, CS^{SG}, and CS^{QA} average 4.3%, 4.6%, and 6.4% below the lower corresponding single-setting score. Strong single-setting results therefore do not ensure correctness under both target specifications.



(a) Performance on OmniCompass-VL (b) Performance on OmniCompass-AV

Figure 4: **Fine-grained QA performance across tracks.** We report category-wise CS^{QA} scores in percent.

OmniCompass-AV. In Table 3, Gemini-3.8-Flash leads in CS^{Avg} at **57.4%**, versus **24.6%** for open-source leader MiMo-V2.5. **(i) Temporal Grounding and Reasoning Remain Challenging.** Qwen3.8-Omni-Flash reaches 73.2% in CS^{QA} but only 21.8% in CS^{TG}; all ten open-source models score at most **3.5%** in CS^{TG}. In Figure 4(b), Temporal Reasoning ranks lowest for five of six models, whereas Causal Reasoning leads all six at 74.3%–94.3%. Qwen3-Omni-30B-A3B scores lowest on Event Understanding at 31.5%. **(ii) Open-Source Models Lag Most in Audio-Visual Event Binding.** In Figure 4(b), the best proprietary and open-source CS^{QA} scores differ by 12.5% on Speech Understanding. The gap reaches 38.9% on Event Understanding, which requires linking audio cues to visible entities, actions, or scene context. **(iii) Reference Effects and Consistency Gaps Persist with Audio.** *Text+Image* improves SG m_{sIoU} and QA accuracy for 12 of 17 models each, but TG m_{tIoU} for only five. MiMo-V2.5 achieves 59.2% and 58.7% QA accuracy under *Text* and *Text+Image*, respectively, but only 47.3% in CS^{QA}. Similar aggregate accuracy therefore does not ensure correct answers on the same eligible pairs across settings.

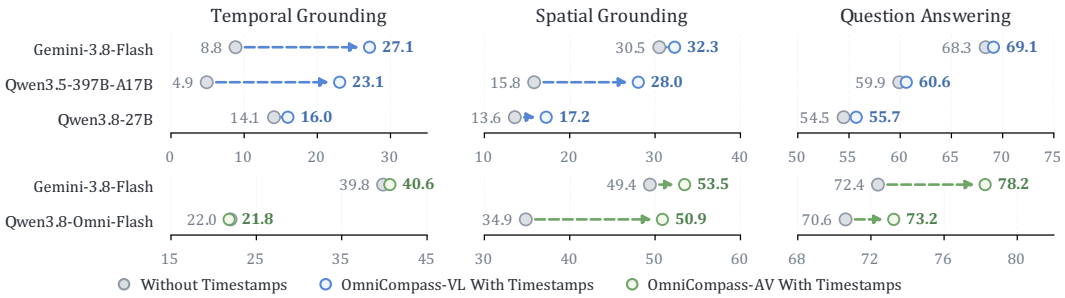


Figure 5: **Impact of frame-level timestamps.** Columns report CS^{TG}, CS^{SG}, and CS^{QA} in percent. Dashed links connect scores without and with timestamps. Axis ranges vary by panel.

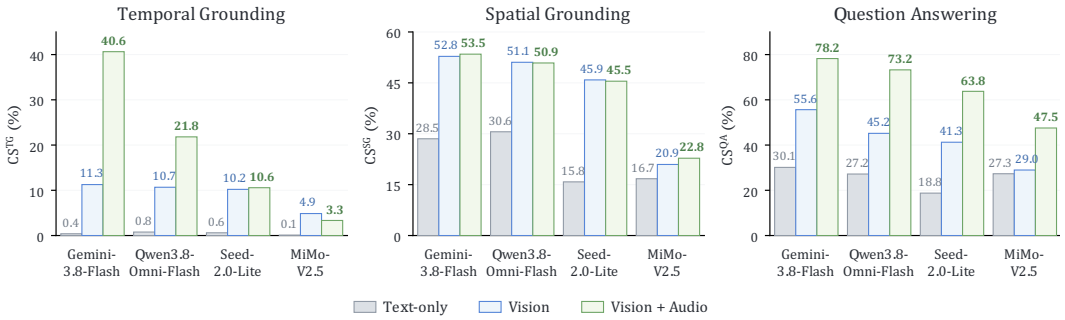


Figure 6: **Impact of input modalities on OmniCompass-AV.** Bars report CS in percent for Text-only, Vision, and Vision+Audio on matched eligible pairs within each model and task.

4.3 Analysis Experiments

Impact of Frame-Level Timestamps. Figure 5 shows that timestamps improve CS^{Avg} in all five model-track configurations, with gains in **14 of 15** task-level comparisons. On the VL track, Gemini-3.8-Flash, Qwen3.5-397B-A17B, and Qwen3.8-27B gain 7.0%, 10.4%, and 2.3%, averaging **6.5%**. On the AV track, Gemini-3.8-Flash and Qwen3.8-Omni-Flash gain 3.6% and 6.2%, averaging **4.9%**. VL gains concentrate in TG: Gemini-3.8-Flash and Qwen3.5-397B-A17B gain **18.3%** and **18.2%**. In AV, Qwen3.8-Omni-Flash gains **16.0%** in SG, while TG changes by -0.2% to +0.8% across both models. QA improves in every comparison by 0.7%–5.9%. These results suggest that timestamps help align sampled frames with temporal positions.

Impact of Input Modalities. Figure 6 compares task-level CS across Text-only, Vision, and Vision+Audio for four OmniCompass-AV models. All comparisons use matched eligible pairs within each model and task. Text-only retains both question wordings but removes all media, including reference images. Its CS^{TG} is only 0.1–0.8%, versus 18.8%–30.1% in CS^{QA}. Vision improves all 12 model-task comparisons. Adding audio raises CS^{QA} by **18.6%–28.1%**, averaging **22.9%**. Gemini-3.8-Flash and Qwen3.8-Omni-Flash gain **29.4%** and **11.1%** in CS^{TG}, respectively, whereas MiMo-V2.5 loses 1.5%. CS^{SG} changes range from -0.4% to +1.8%. Thus, adding audio **consistently improves** CS^{QA} across all four tested models, whereas TG gains vary by model and SG gains remain limited. Additional analyses are provided in Appendix B.

5 Conclusion

In this paper, we introduce OmniCompass, a holistic benchmark for omni-modal video grounding and understanding. It unifies TG, SG, and QA across VL and AV track with paired *Text* and *Text+Image* specifications for the same video, target, task semantics, and ground truth. It covers 820 manually annotated videos across seven domains; CS measures correctness and cross-reference consistency. Leading proprietary models outperform open-source models, especially in audio-visual understanding, while precise grounding, fine-grained vision, and audio-visual binding remain challenging. Reference images yield task-dependent effects, while timestamps improve model performance in both tracks. These results motivate joint evaluation of QA, precise grounding, and cross-reference consistency across tasks and modalities for reliable omni-modal video understanding evaluation.

AI USE STATEMENT

As described in Section 3 and Appendix C, we use `Qwen3-VL-8B-Instruct` to generate detailed video captions and structured descriptions for candidate screening. LLM-based tools also assist with manuscript polishing and figure preparation based on our existing data. The authors review the AI-assisted content and take responsibility for the accuracy and integrity of the paper.

REPRODUCIBILITY STATEMENT

Benchmark construction, annotation protocols, and evaluation metrics are described in Section 3. Experimental setup is provided in Section 4. Details of data filtering and quality verification, along with evaluation prompts and output formats for both tracks, are provided in Appendix C.

REFERENCES

Xiang An, Yin Xie, Feilong Tang, Yunyao Yan, Huajie Tan, Didi Zhu, Changrui Chen, Xiuwei Zhao, Bin Qin, Kaicheng Yang, et al. Llava-onevision-2: Towards next-generation perceptual intelligence. *arXiv preprint arXiv:2605.25979*, 2026.

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, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Jianghan Chao, Jianzhang Gao, Wenhui Tan, Yuchong Sun, Ruihua Song, and Liyun Ru. Jointavbench: A benchmark for joint audio-visual reasoning evaluation. In *Proceedings of the International Conference on Learning Representations*, 2026.

Guo Chen, Yicheng Liu, Yifei Huang, Baoqi Pei, Jilan Xu, Yuping He, Tong Lu, Yali Wang, and Limin Wang. Cg-bench: Clue-grounded question answering benchmark for long video understanding. In *Proceedings of the International Conference on Learning Representations*, 2025a.

Joya Chen, Ziyun Zeng, Yiqi Lin, Wei Li, Zejun Ma, and Mike Zheng Shou. Livecc: Learning video llm with streaming speech transcription at scale. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 29083–29095, 2025b.

Yu Chen, Caorui Li, Ziyu Xiong, Yidong Wang, Mingqi Gao, Shuman Liu, Biao Liu, Chunfeng Yang, Anxiang Zeng, Haibo Zhang, et al. Omnireasoner: Thinking with long audio-video via native tool use. *arXiv preprint arXiv:2607.19339*, 2026.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, Ji Ma, Jiaqi Wang, Xiaoyi Dong, Hang Yan, Hewei Guo, Conghui He, Botian Shi, Zhenjiang Jin, Chao Xu, Bin Wang, Xingjian Wei, Wei Li, Wenjian Zhang, Bo Zhang, Pinlong Cai, Licheng Wen, Xiangchao Yan, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024.

Zixu Cheng, Jian Hu, Ziquan Liu, Chenyang Si, Wei Li, and Shaogang Gong. V-star: Benchmarking video-llms on video spatio-temporal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9155–9164, 2026.

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

Chaoyou Fu, Haozhi Yuan, Yuhao Dong, Yi-Fan Zhang, Yunhang Shen, Xiaoxing Hu, Xueying Li, Jinsen Su, Chengwu Long, Xiaoyao Xie, et al. Video-mme-v2: Towards the next stage in benchmarks for comprehensive video understanding. *arXiv preprint arXiv:2604.05015*, 2026.

Hong Gao, Jingyu Wu, Xiangkai Xu, Kangni Xie, Yunchen Zhang, Bin Zhong, Xurui Gao, and Min-Ling Zhang. Omniground: A comprehensive spatio-temporal grounding benchmark for real-world complex scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17588–17597, 2026.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5267–5275, 2017.

Google. A new era of intelligence with Gemini 3, 2025. URL https://blog.google/products/gemini/gemini-3/.

Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. Worldsense: Evaluating real-world omnimodal understanding for multimodal llms. In *Proceedings of the International Conference on Learning Representations*, 2026.

Wenyi Hong, Yean Cheng, Zhuoyi Yang, Weihan Wang, Lefan Wang, Xiaotao Gu, Shiyu Huang, Yuxiao Dong, and Jie Tang. Motionbench: Benchmarking and improving fine-grained video motion understanding for vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8450–8460, 2025.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, M. C., Jianfeng Cai, Xinyuan Cai, Peizhou Cao, Yuxuan Cao, Ziwei Chai, et al. Kimi K3: Open frontier intelligence. *arXiv preprint arXiv:2607.24653*, 2026.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. Tvqa: Localized, compositional video question answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1369–1379, 2018.

Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. Tvqa+: Spatio-temporal grounding for video question answering. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 8211–8225, 2020.

Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 34, pp. 11846–11858, 2021.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a.

Caorui Li, Yu Chen, Yiyan Ji, Jin Xu, Zhenyu Cui, Shihao Li, Yuanxing Zhang, Jiafu Tang, Zhenghao Song, Dingling Zhang, et al. Omnivideobench: Towards audio-visual understanding evaluation for omni mllms. In *Proceedings of the International Conference on Learning Representations*, 2026.

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

Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Yixuan Li, Lei Chen, Runyu He, Zhenzhi Wang, Gangshan Wu, and Limin Wang. Multisports: A multi-person video dataset of spatio-temporally localized sports actions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13536–13545, 2021.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26286–26296, 2024a.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 8731–8772, 2024b.

David Ma, Yuanxing Zhang, Jincheng Ren, Jarvis Guo, Yifan Yao, Zhenlin Wei, Zhenzhu Yang, Zhongyuan Peng, Boyu Feng, Jun Ma, Xiao Gu, Zhoufutu Wen, King Zhu, Yancheng He, Meng Cao, Shiwen Ni, Jiaheng Liu, Wenhao Huang, Ge Zhang, and Xiaojie Jin. Iv-bench: A benchmark for image-grounded video perception and reasoning in multimodal llms. In *Proceedings of the International Conference on Learning Representations*, 2026.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 36, pp. 46212–46244, 2023.

Jiahao Meng, Xiangtai Li, Haochen Wang, Yue Tan, Tao Zhang, Lingdong Kong, Yunhai Tong, Anran Wang, Zhiyang Teng, Yujing Wang, et al. Open-o3-video: Grounded video reasoning with explicit spatio-temporal evidence. In *Proceedings of the International Conference on Machine Learning*, 2026.

Furkan Mumcu, Michael Jones, Yasin Yilmaz, and Anoop Cherian. Complexvad: Detecting interaction anomalies in video. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pp. 1093–1102, 2025.

OpenAI. GPT-6 Astra: A New Generation of Intelligence, 2026. URL https://openai.com/index/gpt-6-astra/.

Bharathkumar Ramachandra and Michael Jones. Street scene: A new dataset and evaluation protocol for video anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2569–2578, 2020.

Bytedance Seed. Seed2. 0 model card: Towards intelligence frontier for real-world complexity. *arXiv preprint arXiv:2607.00248*, 2026.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18221–18232, 2024.

Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6479–6488, 2018.

Zongheng Tang, Yue Liao, Si Liu, Guanbin Li, Xiaojie Jin, Hongxu Jiang, Qian Yu, and Dong Xu. Human-centric spatio-temporal video grounding with visual transformers. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(12):8238–8249, 2021.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

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

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

Weizhi Wang, Zili Wang, Yuqi Huo, Yuanting Chen, Xuezheng Xu, Jinguo Zhu, Wangbo Zhao, Wei Li, and Weijie Su. Vtool-r1: Vlms learn to think with images via reinforcement learning. In *ICLR*, 2026.

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

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9777–9786, 2021.

Zhenghao Xing, Ruiyang Xu, Yuxuan Wang, Jinzheng He, Ziyang Ma, Qize Yang, Yunfei Chu, Jin Xu, Junyang Lin, Chi-Wing Fu, and Pheng-Ann Heng. Native active perception as reasoning for omni-modal understanding. In *ICML*, 2026.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025a.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025b.

Yudong Yang, Jimin Zhuang, Changli Tang, Guangzhi Sun, Yixuan Li, Peihan Li, Yifan Jiang, Wei Li, Zejun Ma, and Chao Zhang. Audio-centric video understanding benchmark without text shortcut. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2025.

Zuhao Yang, Sudong Wang, Kaichen Zhang, Keming Wu, Sicong Leng, Yifan Zhang, Chengwei Qin, Shijian Lu, Xingxuan Li, and Lidong Bing. Longvt: Incentivizing thinking with long videos via native tool calling. In *CVPR*, 2026.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 9127–9134, 2019.

Xiangyu Zeng, Zhiqiu Zhang, Yuhan Zhu, Xinhao Li, Zikang Wang, Changlian Ma, Qingyu Zhang, Zizheng Huang, Kun Ouyang, Tianxiang Jiang, Ziang Yan, Yi Wang, Hongjie Zhang, Yali Wang, and Limin Wang. Video-o3: Native interleaved clue seeking for long video multi-hop reasoning. In *Proceedings of the International Conference on Machine Learning*, 2026.

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

Jun Zhang, Teng Wang, Yuying Ge, Yixiao Ge, Xinhao Li, and Limin Wang. Timelens: Rethinking video temporal grounding with multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10419–10429, 2026.

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

Zhu Zhang, Zhou Zhao, Yang Zhao, Qi Wang, Huasheng Liu, and Lianli Gao. Where does it exist: Spatio-temporal video grounding for multi-form sentences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10668–10677, 2020.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, et al. Mlvu: Benchmarking multi-task long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13691–13701, 2025a.

Ziwei Zhou, Rui Wang, and Zuxuan Wu. Daily-omni: Towards audio-visual reasoning with temporal alignment across modalities. *arXiv preprint arXiv:2505.17862*, 2025b.

APPENDIX

A Additional Results **15**

A.1 Performance across Video Durations 15

A.2 Grounding across Target Granularity 15

A.3 Performance across Video Domains 16

B Additional Analyses **16**

B.1 Impact of Frame Scaling 16

B.2 Impact of Reference Specification 17

B.3 Cross-Specification Consistency 17

B.4 Qualitative Analysis 18

C Benchmark Construction **20**

C.1 Video Sources and Selection 20

C.2 Task-Oriented Candidate Scoring 22

C.3 Question-Answering Taxonomy 22

C.4 Reference Design and Provenance 22

D Annotation Quality and Verification **23**

D.1 Review Criteria and Repair Records 23

D.2 Human Effort 25

E Evaluation Protocols and Prompts **29**

E.1 Inputs and Paired Settings 29

E.2 Prompt Examples 29

A ADDITIONAL RESULTS

A.1 PERFORMANCE ACROSS VIDEO DURATIONS

Figure A1 compares ten models across short, medium, and long videos in OmniCompass-VL. All ten score lower on long than short videos in TG. Seed-2.1-Pro-260915 shows the largest decline, from 33.4% to 15.1%, while Gemini-3.8-Flash falls from 33.8% to 21.0%. Only five models decline in SG and six in QA. Thus, the direction of the duration association is more consistent for temporal localization than for the other tasks. These groups also differ in content and target granularity, so the comparison does not isolate video duration as the cause of lower scores.

A.2 GROUNDING ACROSS TARGET GRANULARITY

Figure A2 compares ten VL models and six AV models across relative target-duration and box-area groups. All ten VL models improve monotonically in TG as the relative target duration increases. All six AV models also peak in the highest duration-ratio quartile. Gemini-3.8-Flash rises from 7.9% to 43.1% on VL and from 27.1% to 56.2% on AV between the extreme groups. Spatial trends differ: every VL model peaks in the largest-box group, whereas every AV model peaks in the third quartile. Gemini-3.8-Flash scores 65.5% in the third AV quartile and 53.6% in the fourth.

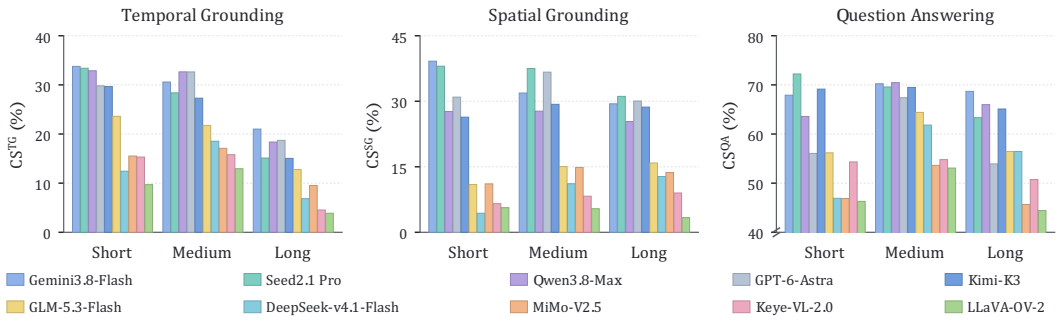
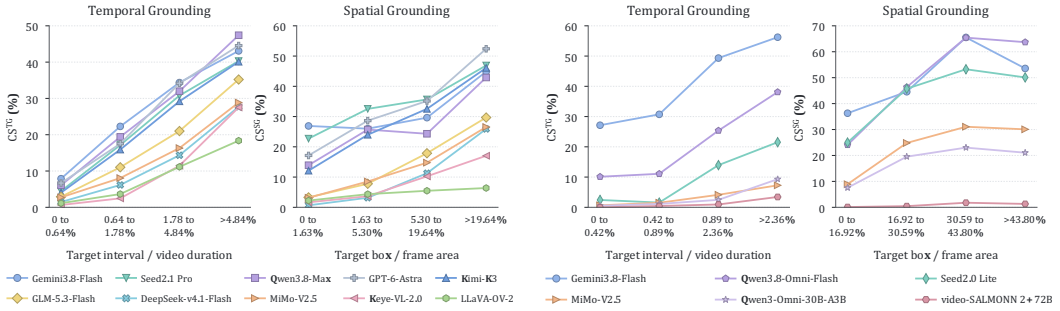


Figure A1: **Performance across video durations in OmniCompass-VL.** Bars show task-specific CS in percent for short videos below 90 seconds, medium videos from 90 to below 300 seconds, and long videos of at least 300 seconds. Each model uses its own eligible pairs.

Quartile boundaries are computed from all videos separately for each track; scores use the common eligible pairs within each task. The resulting groups are not equal in size, and their boundaries differ between VL and AV. IoU also depends on target extent, so these curves describe associations rather than an isolated granularity effect. The between-track contrast does not isolate an audio effect, because the tasks, target distributions, and model rosters also differ.



(a) Grounding Performance on OmniCompass-VL (b) Grounding Performance on OmniCompass-AV

Figure A2: **Grounding across target granularity.** Curves report CS^{TG} and CS^{SG} in percent across quartiles of target-to-video duration and mean endpoint box-to-frame area. Quartiles are track-specific; models share eligible pairs within each track and task.

A.3 Performance across Video Domains

Figure A3 compares the ten VL models across seven domains. All ten obtain their lowest SG and QA scores in Traffic, with equal-weight model means of 7.8% and 37.3%. Variety has the lowest mean TG score at 8.1%, but the highest mean SG score at 26.6%. Strong spatial grounding can therefore coexist with weak temporal localization within the same domain. These rankings apply to the shared eligible subsets, not necessarily every video in each domain. Differences in duration, target size, and question content remain intertwined with the domain labels.

B Additional Analyses

B.1 Impact of Frame Scaling

Increasing the frame budget improves performance in most settings, but the effect depends on the track and task. On OmniCompass-VL, all three tasks improve significantly for both models from 8 to 128 frames. TG CS increases by **22.2%** and 19.6% for Gemini-3.8-Flash and Qwen3.5-397B-A17B, respectively; SG increases by **22.8%** and 16.9%, while QA increases by 20.0% and 18.6%. The gains are smaller and more task-dependent on OmniCompass-AV. SG improves by 18.3% for Gemini-3.8-Flash and **22.3%** for Qwen3.8-Omni-Flash, while QA improves by 9.9% and 13.4% from 8 to 128 frames. In contrast, TG remains nearly unchanged. We suppose that audio already provides useful temporal cues for AV TG (see Figure 6), leaving denser visual sampling with limited

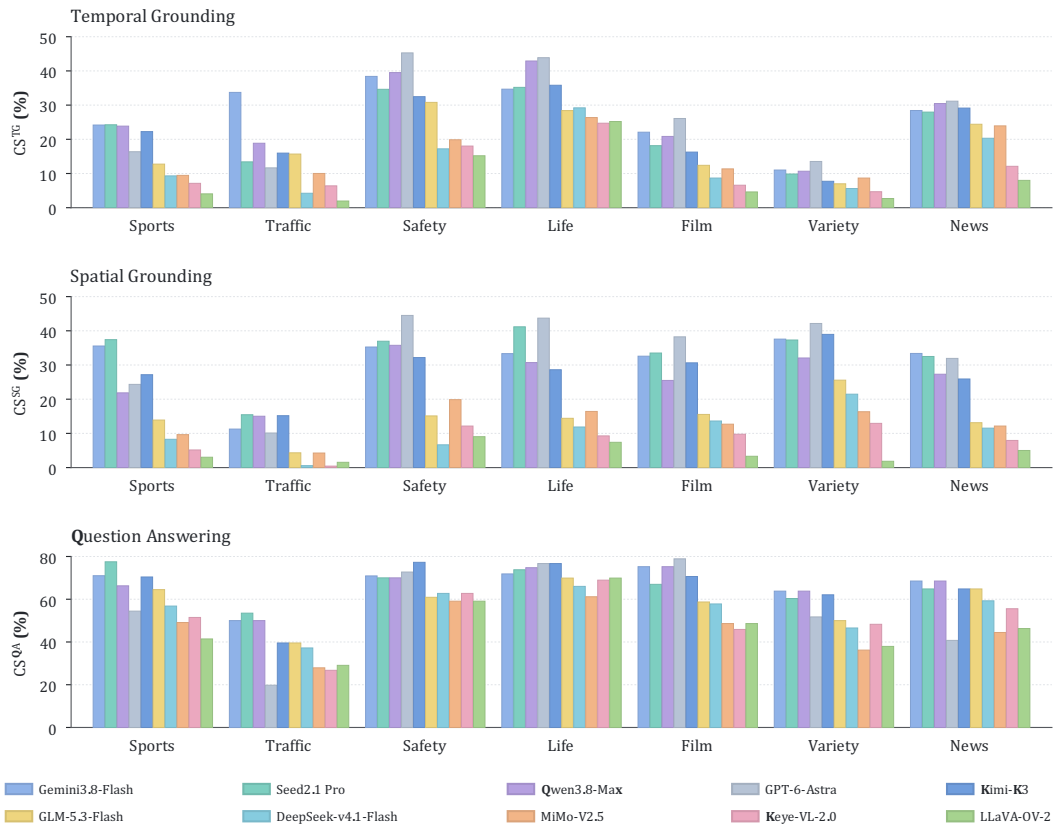


Figure A3: **Performance across video domains in OmniCompass-VL.** Bars report task-specific CS in percent for ten models. The shared eligible subsets contain 592 TG, 673 SG, and 689 QA pairs across the seven domains; comparisons use the same instances within each task.

additional temporal evidence, whereas SG and QA can benefit from finer spatial details and richer visual context. Besides, more frames also increase visual input and inference cost, so the preferred budget depends on the task and track rather than always favoring the largest setting.

B.2 Impact of Reference Specification

Figure B2 compares *Text* and *Text+Image* across reference types. For action-conditioned questions, *Text+Image* improves TG for seven of ten models and SG for nine, but lowers QA accuracy for eight. Gemini-3.8-Flash gains 10.4% in action-conditioned SG, whereas Qwen3.8-Max loses 7.6%. The paired settings change the target description as well as image availability, so these differences are not a pure image-addition effect. They measure whether performance changes when a visual cue replaces part of the textual target description under matched evidence and semantics.

Each task uses the same eligible subset across all ten models. Action-conditioned TG, SG, and QA contain 50, 54, and 31 pairs, respectively. Scene-conditioned TG contains only eight pairs, so its cell values provide limited evidence for a general trend. The heatmap reports mean differences, not significance; subgroup sizes matter when interpreting their direction.

B.3 Cross-Specification Consistency

Figure B3 compares paired CS with setting-specific performance. Across the ten models, TG and SG CS average 5.0% and 5.7% below the lower single-setting mean. This gap includes both the cost of requiring overlap with ground truth under both settings and the additional prediction-to-prediction overlap constraint. The latter contributes only 0.3%–1.8% across the model–task combinations. The full gap does not measure prediction disagreement alone.

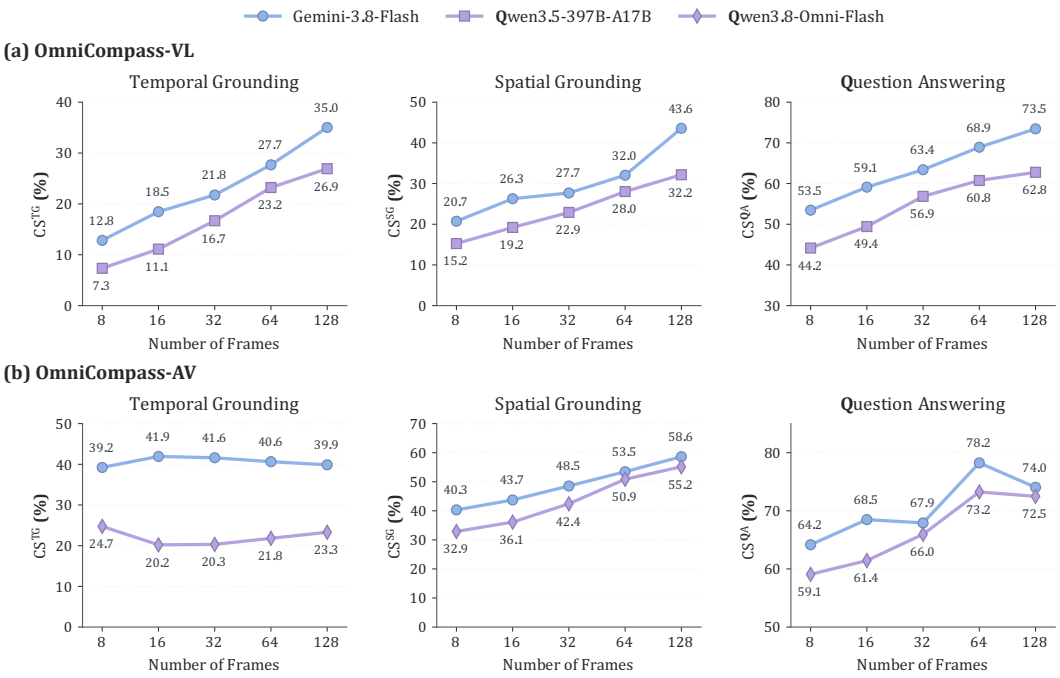


Figure B1: **Impact of frame scaling.** Each panel reports task-level paired CS (%) under 8, 16, 32, 64, and 128 uniformly sampled frames. Rows correspond to the VL and AV tracks, columns correspond to TG, SG, and QA, and curves distinguish the evaluated models.

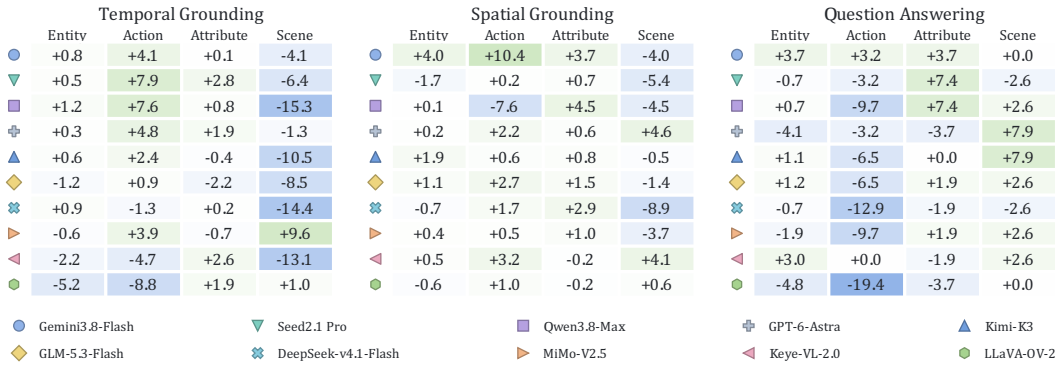


Figure B2: **Effects of reference specification in OmniCompass-VL.** Cells show *Text+Image* minus *Text* in percent for TG m_{tIoU}, SG m_{sIoU}, and QA accuracy on shared per-task subsets. Green denotes increases and blue denotes decreases; color does not encode significance.

For QA, 10.5–25.1% of eligible pairs are answered correctly under exactly one setting. Kimi-K3 reaches nearly identical accuracies of 73.0% and 72.9%, yet answers only 67.6% correctly under both. Its 808 pairs comprise 546 correct under both settings, 44 only under *Text*, 43 only under *Text+Image*, and 175 under neither. Similar aggregate accuracy can therefore mask different instance-level outcomes. Neither-correct pairs may still contain two different wrong answers.

B.4 Qualitative Analysis

As shown in Figure B4, we provide qualitative examples of OmniCompass-VL and OmniCompass-AV across QA, SG, and TG tasks. In the first OmniCompass-VL case, the older man drinks twice, corresponding to option C. Seed-2.1-Pro predicts C under both *Text* and *Text+Image*, while Kimi-K3 changes from A to C and GPT-6-Astra changes from A to B after the reference image is provided, rather than text only. In the second OmniCompass-VL case, the gymnastics team ends in the formation described by option C, with three members kneeling on one knee and three standing. Seed-2.1-Pro and Gemini-3.8-Flash select C, whereas GPT-6-Astra selects A.

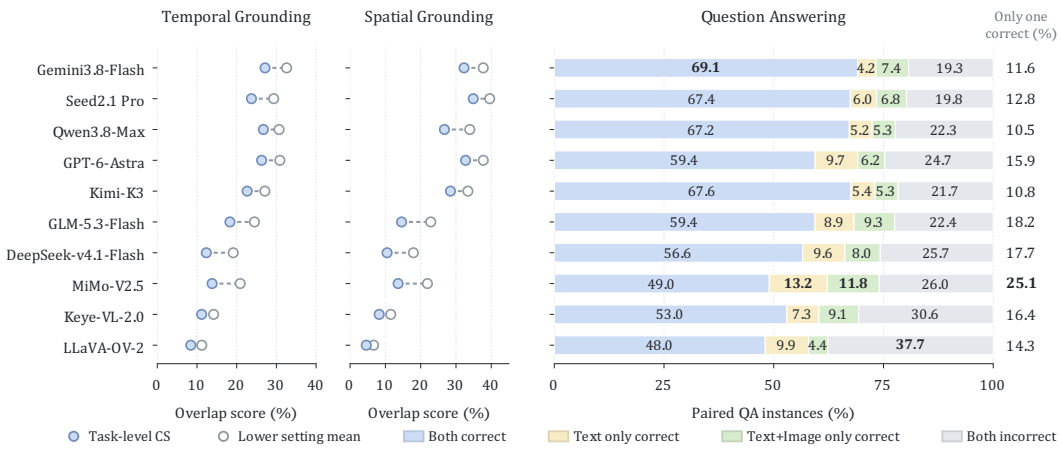


Figure B3: **Consistency under paired reference specifications.** Grounding compares CS with the lower setting-specific mean on identical eligible pairs. QA bars partition pairs by correctness in the two settings; blue denotes CS^{QA}. The right column reports correctness under exactly one setting. Bold values mark category maxima, not a uniform preference for higher values.

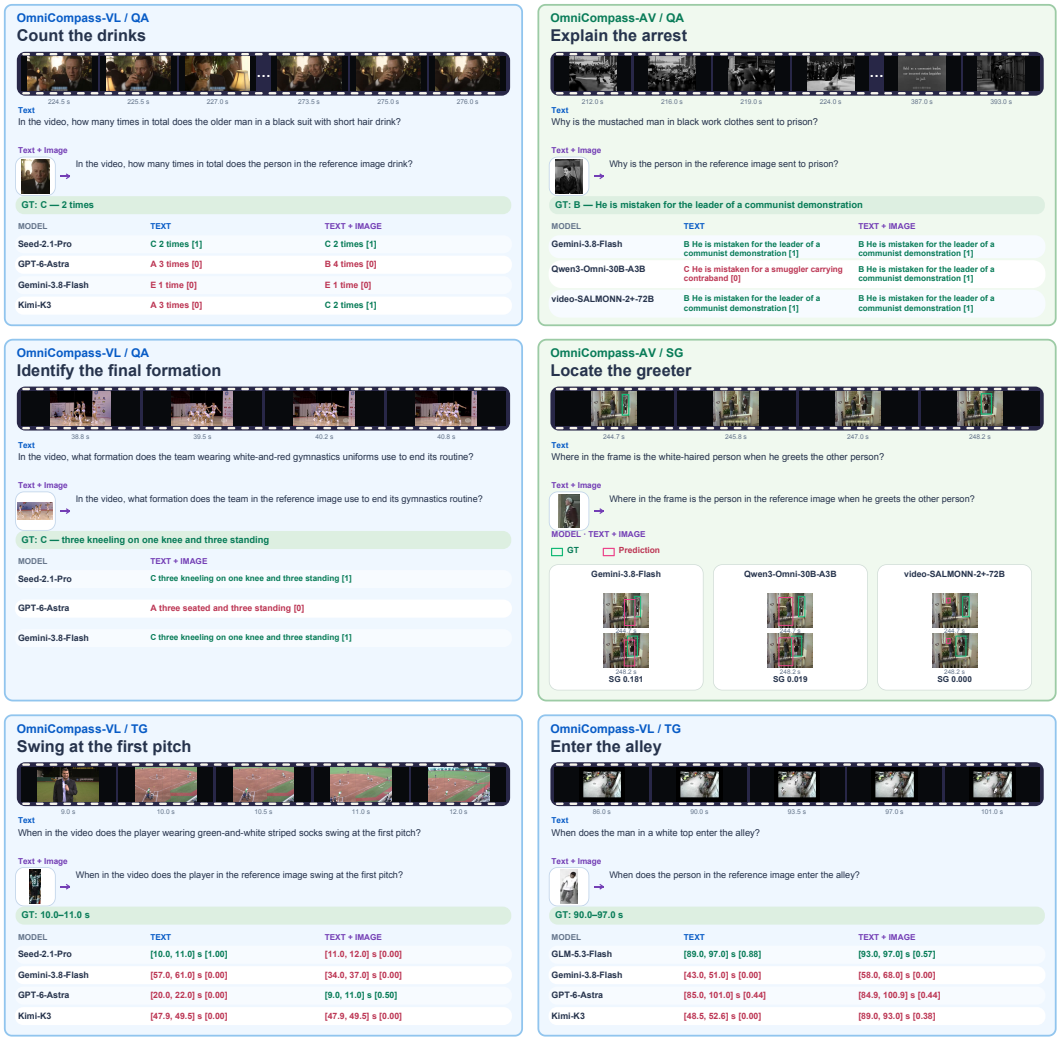


Figure B4: **Qualitative Analysis on our OmniCompass-VL and OmniCompass-AV.** Different models predictions on six cases in our two evaluation tracks.

For OmniCompass-AV, the upper-right QA example contains an intertitle indicating that the man is mistaken for the leader of a communist demonstration, corresponding to option B. Qwen3-Omni-30B-A3B changes from C under *Text* to B under *Text+Image*, while Gemini-3.8-Flash and video-SALMONN-2+-72B select B in both settings. The lower-right SG example illustrates the difficulty of SG: Gemini-3.8-Flash obtains the highest score among the three models, but its score is only 0.181, compared with 0.019 for Qwen3-Omni-30B-A3B and 0.000 for video-SALMONN-2+-72B.

The last two OmniCompass-VL TG examples highlight variation in temporal localization. In the baseball example, the annotated interval for the first-pitch swing is 10.0–11.0 s. Seed-2.1-Pro predicts this interval exactly under *Text* but shifts to 11.0–12.0 s under *Text+Image*; GPT-6-Astra moves from 20.0–22.0 s to 9.0–11.0 s, yielding a partial overlap with the annotated interval. In the CCTV example, the annotated entrance interval is 90.0–97.0 s. GLM-5.3-Flash achieves the highest score under *Text* with the prediction 89.0–97.0 s, whereas Kimi-K3 improves from 48.5–52.6 s to 89.0–93.0 s when the reference image is added. Gemini-3.8-Flash remains far from the annotated interval in both settings, and GPT-6-Astra produces a broad interval covering 85.0–101.0 s. These examples illustrate distinct failure modes and case-specific changes across tasks and input settings.

C Benchmark Construction

C.1 Video Sources and Selection

OmniCompass-VL contains 580 selected public-source videos and 240 self-collected clips. The public component draws from nine dataset families; Table C1 reports the contribution of each source to the 820-video set. The self-collected component comprises 70 movie clips, 70 TV-series clips, 50 variety-show clips, and 50 news clips. These same 240 videos also support OmniCompass-AV, with separate audio-visual task annotations. The two tracks share videos, but their TG, SG, and QA annotations may select different events and targets from that shared video content. Figure C1 provides an overview of the seven high-level video categories in our OmniCompass.

Table C1: **Nine public source datasets contributing to OmniCompass.** Counts refer to selected videos, not complete source releases. Domains follow the final OmniCompass taxonomy; News includes documentary, nature, and knowledge content. Links point to downloads.

Source	Videos	Selected domains	License / Terms	Link
`MultiSports` (Li et al., 2021)	83	Sports	CC BY-NC 4.0	`https://huggingface.co/datasets/MCG-NJU/SportsAction`
`LiveSports-3K` (Chen et al., 2025b)	92	Sports	Not specified	`https://huggingface.co/datasets/stdKonjac/LiveSports-3K`
`UCF-Crime` (Sultani et al., 2018)	111	Safety	CC0-1.0 on mirror	`https://huggingface.co/datasets/AllenXeon/ucf_crime`
`CG-Bench` (Chen et al., 2025a)	58	Film, Variety, News, Life, Safety, Traffic	MIT; videos for academic research only	`https://huggingface.co/datasets/CG-Bench/CG-Bench`
`MovieChat-1K-test` (Song et al., 2024)	20	News	Not specified	`https://huggingface.co/datasets/Enxin/MovieChat-1K-test`
`ActivityNetQA` (Yu et al., 2019)	37	Life	MIT; non-commercial use	`https://huggingface.co/datasets/lmms-lab/ActivityNetQA`
`QVHighlights` (Lei et al., 2021)	88	Life, News	CC BY-NC-SA 4.0 for annotations	`https://huggingface.co/datasets/yeliudev/VideoMind-Dataset/tree/main/qvhighlights`
`StreetScene` (Ramachandra & Jones, 2020)	32	Traffic	CC BY-SA 4.0	`https://zenodo.org/records/10870472`
`ComplexVAD` (Mumcu et al., 2025)	59	Traffic	CC BY-SA 4.0	`https://zenodo.org/records/15707073`



Figure C1: **Seven categories of videos in our OmniCompass.**

Table C1 documents provenance rather than a blanket license for the underlying media. Our annotations and the source videos are distinct assets; reuse of a dataset repository does not itself establish permission to redistribute its videos. The self-collected clips are used for academic evaluation, with rights to the original programs retained by their owners. Access and redistribution must respect the applicable source terms, including any restrictions on the videos and reference images.

We use `Qwen3-VL-8B-Instruct` (Bai et al., 2025a) to profile 8,489 domain-aligned public candidates. The outputs describe entities, actions, attributes, and scenes, together with a video description and corresponding counts. The archived records use `identity` for entity information. Figure C2 presents the profiling template with this field displayed as `entity` for consistency with the paper's terminology. These automatic descriptions guide candidate selection; annotators construct questions, localize the required evidence, and verify answers.

Deterministic filtering removes 1,454 candidates shorter than 30 seconds, leaving 7,035 videos. A further 89 lack at least one of the four cue types, yielding 6,946 eligible public candidates. The filtering criteria require positive entity, action, attribute, and scene counts alongside the metadata

Structured Video Profiling Prompt

You are a factual video captioning and structured extraction assistant.
Only describe what is clearly visible or audible. Do NOT guess.
Output ONE valid JSON object only.
Do NOT place JSON inside any field. **Write:**

1. `detailed_description`: 3–6 sentences, time order.
2. `entity`: list categories with `instance_count` and short notes.
3. `action`: up to 12 distinct temporally localizable action variants.
4. `attribute`: up to 12 distinct visible cues.
5. `scene`: up to 6 distinct phases, shots, or contexts.
6. `counts`: `entity` = sum of `entity.instance_count`; others = list lengths.

Required JSON Output

```
Return JSON with EXACT keys:
{
"detailed_description":"...",
"entity":[{"category":"...","instance_count":0,"notes":"..."}],
"action":["..."],
"attribute":["..."],
"scene":["..."],
"counts":{"entity":0,"action":0,"attribute":0,"scene":0}
}
```

Figure C2: **Structured video profiling prompt used with `Qwen3-VL-8B-Instruct`.**

used by the selection pipeline. These candidate-pool counts precede manual task construction and should not be interpreted as counts of valid benchmark questions.

C.2 Task-Oriented Candidate Scoring

Candidate ranking combines reference necessity R, grounding suitability G, temporal-spatial utility T, and evidence diversity E as $S = 0.35R+0.25G+0.25T+0.15E$. These heuristic components use structured profiling statistics, not model evaluation scores. Table C2 defines the recorded component calculations on a zero-to-ten scale. Scores inform shortlisting together with domain, duration, source, and subcategory coverage; human curation determines the retained public set.

C.3 Question-Answering Taxonomy

QA types are grouped by the dominant operation needed to answer each question. Perception covers direct recognition and judgments from the available evidence, while reasoning covers temporal, causal, or comparative relations. Both tracks use five answer options, but their type sets differ because AV questions are designed to draw on speech, sound, or joint audio-visual evidence. The shared format standardizes scoring without equating difficulty across tracks.

OmniCompass-VL contains eight QA types, listed in Table C3 and shown in Figure C3. Fact Recognition, Action Recognition, Attribute Perception, Visual Counting, and Spatial Perception form the perception group. Temporal Reasoning, Causal Reasoning, and Comparative Judgment form the reasoning group. Labels follow the answering operation, not individual words in the question.

OmniCompass-AV contains five QA types, listed in Table C4 and shown in Figure C3. Speech Understanding, Event Understanding, and Emotion Recognition form the perception group; Temporal Reasoning and Causal Reasoning form the reasoning group. Event Understanding links audible cues to visible participants, actions, states, or context. These labels describe intended evidence requirements, rather than a measured guarantee that every item fails when audio is removed.

C.4 Reference Design and Provenance

Reference images identify a target through entity, action, attribute, or scene information, as defined in Table C5. The unit of analysis is a task-specific reference, not a unique image file or video. Each

Table C2: **Mapping from structured video profiling outputs to task-oriented scoring signals.** We summarize how the structured video profiling outputs produced by `Qwen3-VL-8B-Instruct` are converted into intermediate statistics and then used to compute the four score components.

Component	Derived profiling fields	Computation	Intuition
Reference Necessity (R)	`entity_count`, `action_count`, `attribute_count`, `scene_count`, repeated actions	We first compute *ambiguity* and *richness*. Ambiguity increases with multiple candidate entities, repeated actions, scene variation, and potential attribute confusion. Richness increases with broader coverage over entity, action, attribute, and scene signals. We then combine them as $R = 10 \cdot (0.60A + 0.20B + 0.20AB)$, where A and B are the normalized ambiguity and richness scores.	Measures whether external referential cues are needed to uniquely specify the target.
Grounding Suitability (G)	`entity_count`, `action_count`, `attribute_count`, `scene_count`	We estimate temporal grounding potential from action and scene counts, spatial grounding potential from entity and attribute counts, and question answering potential from action and scene-related complexity. The final score is $G = 10 \cdot \text{avg}(P_{\text{TG}}, P_{\text{SG}}, P_{\text{QA}})$.	Measures whether the video can support constructing TG, SG, and VideoQA instances.
Temporal-Spatial Utility (T)	video duration, `entity_count`, `action_count`, `scene_count`	We combine a duration term and a variation term: $T = 10 \cdot (0.60D + 0.40V)$, where D is a log-normalized duration score and V is the average of normalized action, scene, and entity variation.	Measures whether the video provides sufficient temporal span and visual variation for evidence localization.
Evidence Diversity (E)	`entity_count`, `action_count`, `attribute_count`, `scene_count`	We compute normalized entropy over the four evidence counts: $E = 10 \cdot H_{\text{norm}}(c_{\text{ent}}, c_{\text{act}}, c_{\text{att}}, c_{\text{scn}})$.	Measures whether referential evidence is distributed across multiple cue types rather than concentrated in a single source.
Final Score (S)	R, G, T, E	$S = 0.35R + 0.25G + 0.25T + 0.15E$	Encourages videos that require diverse referential cues, support TG/SG/VideoQA construction, and provide diverse evidence for annotation.

of the 820 VL videos contributes one reference for TG, SG, and QA, yielding 2,460 task instances. Tasks can select different targets; their cue types and sources are counted separately.

Entity cues account for 1,960 of the 2,460 references, or 79.7%. The remaining references comprise 171 action cues, 200 attribute cues, and 129 scene cues. The examples include recognizable actions in Sports and appearance-based vehicle distinctions in Traffic. Figure C4 reports domain-wise frequencies; the cue types are not equally represented in this set.

The reference set contains 1,933 video-internal images and 527 Internet images. Video-internal crops preserve the identity of targets specific to a recording, while external exemplars can specify recognizable people, objects, or actions. Source selection follows the identification needs of each annotation rather than a fixed per-domain quota. This variation is part of the benchmark's reference conditions and is not an independently controlled source-modality experiment.

For video-internal references, annotation guidelines favor a separate view outside the target interval or QA clue interval whenever one is available. Reviewers also inspect whether the image reveals the answer event, since temporal separation alone does not rule out semantic leakage. Crops retain target-identifying content while limiting irrelevant context and direct answer evidence. These checks aim to separate target specification from the evidence needed to solve the task.

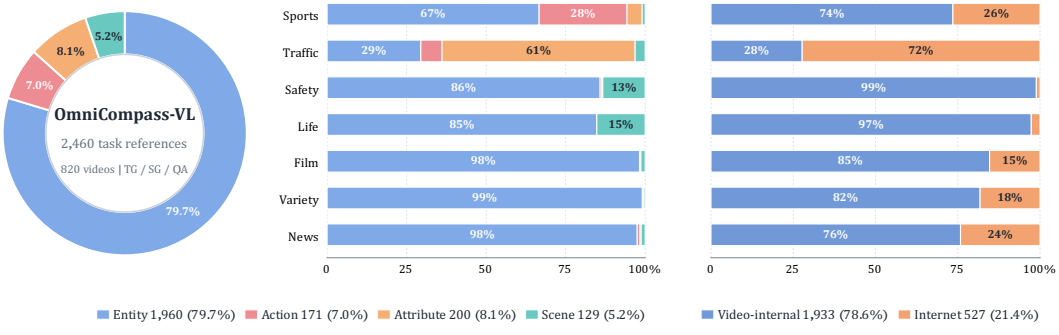
D Annotation Quality and Verification

D.1 Review Criteria and Repair Records

Quality control combines semantic review with structural validation. Reviewers check paired-setting target identity, temporal-event completeness, endpoint box alignment, QA answer uniqueness, and evidence sufficiency. AV review additionally examines speech or sound dependence and audio-visual correspondence. Returned corrections undergo task-level re-verification and lead-expert acceptance. Tables D1 and D2 summarize verification stages and representative issues.



Figure C3: **QA taxonomy in our OmniCompass-VL and OmniCompass-AV.**



(a) Overall Reference Types (b) Reference Types by Domain (c) Image Sources by Domain

Figure C4: **Reference composition across 2,460 VL task instances.** Panel **(a)** shows overall cue-type frequencies, **(b)** shows their proportions by domain, and **(c)** compares video-internal and Internet sources. Counts are task-level, including reused images.

Table C3: **QA taxonomy and examples in OmniCompass-VL.** We group the eight QA types into Perception and Reasoning. Examples show only the question text from the *Text* target-specification setting; answer options are omitted.

Category	QA Task	Task Definition	Example
Perception	Fact Recognition	Identify entities or explicit facts associated with the referenced target.	In the video, when the man in a black suit sitting on the left side of the table is talking to two other people, what is the drink in front of the woman?
	Action Recognition	Identify an action performed by the referenced target.	In the video, what farewell gesture did the young man in the pink hoodie make to each of the two women inside the house before leaving?
	Attribute Perception	Identify the target's appearance, attributes, or observable state.	In the video, when a lady wearing a red-brimmed straw hat dances with a group of people in yellow clothes, what is the color of the oil-paper umbrella held by the lady behind her?
	Visual Counting	Count referenced entities, actions, or event occurrences.	In the video, how many rotations did the athlete complete while raising one hand and high-fiving a partner?
	Spatial Perception	Determine scene context, locations, or spatial relations involving the referenced target.	In the video, where is the kettle relative to the man in a gray hat and light blue short-sleeved shirt, based on the direction he is facing while making tea?
Reasoning	Temporal Reasoning	Identify events or states before, during, or after a specific referenced event.	Before the man in a blue shirt enters the first elevator, what type of passenger is among the previous group entering the same elevator?
	Causal Reasoning	Infer causes, effects, or motives that connect observed events.	In the video, why does the boy with short blond hair repeatedly throw away the eggs in secret rather than return or refuse them directly?
	Comparative Judgment	Compare targets, events, states, or outcomes to determine differences, rankings, or changes.	In the video, when several people on stage take turns answering the question asked by the female host wearing a colorful patterned sleeveless top, what is the relationship among their answering durations?

D.2 Human Effort

The construction and verification records yield a retrospective estimate of 1,784 human-hours, or 223 eight-hour person-days. This comprises 1,280 annotator-hours and 504 lead-expert-hours, as detailed in Table D3. The estimate covers training, task construction, cross-verification, correction, adjudication, and final auditing. It is reconstructed from project records rather than continuous time logs and excludes automated processing, model inference, and waiting periods.

Table C4: **QA taxonomy and examples in OmniCompass-AV.** We group the five QA types into Perception and Reasoning. Examples show only the question text from the *Text* target-specification setting; answer options are omitted.

Category	QA Task	Task Definition	Example
Perception	Speech Understanding	Understand explicit speech content, including facts, quantities, descriptions, and opinions.	How does the gray-haired man in a black leather jacket and white top summarize what the Nike swoosh represented to young people at the time?
	Event Understanding	Identify participants, objects, actions, states, or counts in audio-visual events.	When the woman with tied-up hair, a blue off-the-shoulder top, and brown overalls introduces Sean's recently deceased father, which pair of objects appears beside the relevant framed photograph?
	Emotion Recognition	Recognize emotion, tone, attitude, or affective reactions from audio-visual evidence.	In the video, after repeated corrections from the man in red, how does the woman with shoulder-length auburn hair sound when asking whether he wants to tell the story?
Reasoning	Temporal Reasoning	Determine the order or temporal relations of utterances, events, and states.	In the video, before the blonde woman calls the grey bearded man arrogant, what is he doing?
	Causal Reasoning	Infer causes, motives, or consequences from audio-visual context.	Why could the lie told by the woman in a red choker and black top to avoid awkwardness with Joey end up costing her job?

Table C5: **Functional taxonomy of reference information in OmniCompass-VL.** Each type states what the reference image contributes to target identification and gives representative examples.

Reference Type	Role in Target Identification and Typical Examples
`Entity`	Shows a particular person or object that the model must match across views, such as an athlete, actor, vehicle, or bag.
`Action`	Shows a recognizable motion or event pattern that defines the target behavior, such as a dunk, basketball shot, header, or tackle.
`Attribute`	Distinguishes the target through a visible property, appearance, or state, such as clothing color, vehicle type, shape, or pose.
`Scene`	Shows a place or surrounding visual context to match with a video segment, such as a street, room, stage, or stadium.

Table D1: **Multi-stage quality verification for OmniCompass-VL.** All counts use the final 820-video release set. A video passes only when its paired *Text* and *Text+Image* specifications and its TG, SG, and QA annotations satisfy the verification criteria. Representative findings are non-exclusive.

Verification stage	Videos	Stage outcome	Representative findings	Cumulative pass
Annotator cross-verification	820	466 require no revision; 354 are returned for revision.	Cross-setting target mismatch; a *Text+Image* pointer identifiable without its image; imprecise TG boundaries; incorrect or incomplete SG targets; insufficient QA evidence, non-unique answers, or weak distractors.	**466/820** (**56.8%**)
Annotator revision and lead-expert acceptance	354	All revised cases pass after annotator recheck and lead-expert acceptance.	Typical repairs address overly small or partial SG targets, TG intervals that omit part of the event, and inconsistencies among QA questions, answers, types, and evidence.	**820/820** (**100%**)
Final release audit	820	All VL cases are included only after lead-expert acceptance.	Same target and ground truth across settings; complete TG events; valid boundary-frame boxes; one evidence-supported QA answer with plausible distractors.	**820/820** (**100%**)

Table D2: **Multi-stage quality verification for OmniCompass-AV.** All counts use the 240-video AV subset of the final release. In addition to the shared VL criteria, annotators check native-audio dependence and audio-visual alignment, and the lead expert includes these criteria in final acceptance and audit. Representative findings are non-exclusive.

Verification stage	Videos	Stage outcome	Representative findings	Cumulative pass
Annotator cross-verification (AV-specific checks)	240	151 required no targeted repair; 89 entered repair for 105 task-level issues.	Insufficient audio dependence; clue intervals omitting required speech or sound; audio-visual target mismatch; SG boxes following the wrong person; unsupported QA answers.	**151/240** (**62.9%**)
Annotator revision and lead-expert acceptance	89	24 TG, 70 SG, and 11 QA annotations corrected; all 89 videos accepted.	TG intervals starting too early; SG references, questions, and boxes identifying different targets; QA type or evidence mismatch; imbalanced distractors.	**240/240** (**100%**)
Final release audit	240	All AV cases are included only after lead-expert acceptance and audit.	Required native-audio evidence; aligned audio-visual clues; the same target and answer across settings; valid TG, SG, and QA annotations.	**240/240** (**100%**)

Table D3: **Estimated human effort for OmniCompass construction and quality assurance.** Hours are reconstructed from dated project records using an eight-hour workday. Automated processing, model inference, and waiting periods are excluded.

Stage	Personnel	Scope	Person-days	Hours
Guideline training	4 annotators	Two-day guideline training and pilot-protocol calibration.	8	64
Initial task construction	4 annotators	Paired *Text* and *Text+Image* TG, SG, and QA instances for 820 VL videos, including the 240-video AV subset.	112	896
Mutual cross-verification and targeted revision	4 annotators	Mutual cross-verification and corrections within the 354 flagged VL cases and track-specific AV repairs identified during verification.	40	320
Final acceptance and correction	Lead expert	Final acceptance of annotator-confirmed cases, with remaining corrections for target pairing, reference necessity, TG/SG validity, QA evidence, and AV dependence.	31	248
Repair adjudication	Lead expert	Repair-package review, disputed-case resolution, and re-verification of corrected annotations.	16	128
Lead-expert final release audit	Lead expert	Paired-setting consistency, dashboard readback, residual-error checks, and final release validation.	16	128
Total	**4 annotators + 1 lead expert**	**End-to-end human construction and quality assurance.**	**223**	**1,784**

E Evaluation Protocols and Prompts

E.1 Inputs and Paired Settings

Frames are sampled uniformly across full videos. The standard budget is 64 for proprietary models and open-source models above 100B total parameters, and 32 otherwise. GPT-series runs use 49 frames to reserve a reference slot within the evaluated 50-image input limit; VITA-1.5-7B uses 16. Budgets remain fixed within a model across tasks and reference settings. Cross-model results therefore reflect the stated input budgets, not a strictly equal-compute comparison.

For the 52-second VL clip in Figure E1, the 64-frame inputs are:

Text:
```
Timestamp: 0 seconds. <frame 0> →
Timestamp: 0.84 seconds. <frame 1> →
...
Timestamp: 51.96 seconds. <frame 63> →
[Prompt]
```

Text+Image:
```
Timestamp: 0 seconds. <frame 0> →
Timestamp: 0.84 seconds. <frame 1> →
...
Timestamp: 51.96 seconds. <frame 63> →
<ref image> → [Prompt]
```

For the 270-second AV clip in Figure E4, audio follows the sampled frames:

Text:
```
Timestamp: 0 seconds. <frame 0> →
Timestamp: 4.3 seconds. <frame 1> →
...
Timestamp: 269.967 seconds. <frame 63> →
<audio> → [Prompt]
```

Text+Image:
```
Timestamp: 0 seconds. <frame 0> →
Timestamp: 4.3 seconds. <frame 1> →
...
Timestamp: 269.967 seconds. <frame 63> →
<audio> → <ref image> → [Prompt]
```

Each timestamp is literal text in seconds from the start of the video, placed immediately before its corresponding frame, following the design of TimeLens (Zhang et al., 2026). The frame numbers index sampled frames from zero: `<frame 0>` is the first, and `<frame 63>` is the last in these 64-frame examples. Angle-bracketed items denote media content, not literal text tokens. `<audio>` is the video's audio track, and `<ref image>` is the task-specific reference image. The arrows indicate input order, and `...` omits intermediate timestamp–frame pairs; neither is sent to the model. `[Prompt]` stands for the task prompt, which includes the full video duration. The timestamps shown are specific to these clips, not fixed for other videos or frame budgets. Within each pair, the video evidence, task semantics, and ground truth are shared; only the target specification changes.

TG returns temporal boundaries, and QA selects one of five answer options. Following V-STaR (Cheng et al., 2026), SG receives the relevant interval and returns one bounding box at each endpoint. It therefore evaluates target localization at supplied times, not temporal search or dense tracking throughout the interval. Figures E1–E6 illustrate response formats with annotated ground truth. The output panels are never supplied to the models in either setting.

E.2 Prompt Examples

Figures E1–E3 and E4–E6 show OmniCompass-VL and OmniCompass-AV prompts, respectively.

Text

Input



Video frames

Prompt

You are given a sequence of frames uniformly sampled across a 52-second video. Each video frame is preceded by its timestamp in seconds. Complete the task based on the visual content of these video frames.

Question: When in the video did Team USA player No. 7 score her first basket?

Output format: Answer with only this JSON object, where `start_time` and `end_time` are measured in seconds from the beginning of the video:

```
{"start_time": <number>,
"end_time": <number>}
```

Output

Ground Truth: `{"start_time": 48.0, "end_time": 50.0}`

Text+Image

Input



Video frames

+



Reference image

Prompt

You are given a sequence of frames uniformly sampled across a 52-second video, followed by a reference image. Each video frame is preceded by its timestamp in seconds. Complete the task based on the visual content of the video frames and the reference image.

Question: When in the video did the player in the reference image score her first basket?

Output format: Answer with only this JSON object, where `start_time` and `end_time` are measured in seconds from the beginning of the video:

```
{"start_time": <number>,
"end_time": <number>}
```

Output

Ground Truth: `{"start_time": 48.0, "end_time": 50.0}`

Figure E1: **Temporal Grounding prompts for OmniCompass-VL.** Both settings receive the same uniformly sampled, timestamped video frames without audio and predict the same event interval. Only the target specification changes from a textual description to a reference image.

Text

Input



Video frames

Prompt

You are given a sequence of frames uniformly sampled across a 52-second video. Each video frame is preceded by its timestamp in seconds. Complete the task based on the visual content of these video frames.

Question: Where in the frame is Team USA player No. 7 from when she begins her shooting motion until the ball goes in?

The relevant interval starts at 48.72 seconds and ends at 50.12 seconds. Locate the target described in the question at each of these two timestamps and report one bounding box for each frame.
Write each box in `[x1, y1, x2, y2]` order, corresponding to `[left, top, right, bottom]`. Normalize all coordinates from 0 to 1000, with the top-left of the frame as the origin.

Output format: Return only this JSON object:

```
{"start_time": 48.72,
"start_bbox": [<x1>, <y1>,
<x2>, <y2>],
"end_time": 50.12,
"end_bbox": [<x1>, <y1>, <x2>,
<y2>]}
```

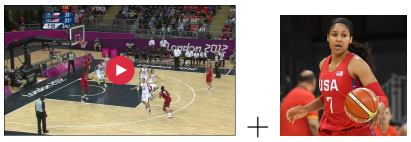
Output

Ground Truth:

```
{"start_time": 48.72,
"start_bbox": [587.61, 473.21, 628.53,
666.34],
"end_time": 50.12,
"end_bbox": [644.16, 536.71, 694.75,
833.00]}
```

Text+Image

Input



Video frames + Reference image

Prompt

You are given a sequence of frames uniformly sampled across a 52-second video, followed by a reference image. Each video frame is preceded by its timestamp in seconds. Complete the task based on the visual content of the video frames and the reference image.

Question: Where in the frame is the player in the reference image from when she begins her shooting motion until the ball goes in?

The relevant interval starts at 48.72 seconds and ends at 50.12 seconds. Locate the target described in the question at each of these two timestamps and report one bounding box for each frame.
Write each box in `[x1, y1, x2, y2]` order, corresponding to `[left, top, right, bottom]`. Normalize all coordinates from 0 to 1000, with the top-left of the frame as the origin.

Output format: Return only this JSON object:

```
{"start_time": 48.72,
"start_bbox": [<x1>, <y1>,
<x2>, <y2>],
"end_time": 50.12,
"end_bbox": [<x1>, <y1>, <x2>,
<y2>]}
```

Output

Ground Truth:

```
{"start_time": 48.72,
"start_bbox": [587.61, 473.21, 628.53,
666.34],
"end_time": 50.12,
"end_bbox": [644.16, 536.71, 694.75,
833.00]}
```

Figure E2: **Spatial Grounding prompts for OmniCompass-VL.** Both settings receive the same uniformly sampled, timestamped video frames and the same temporal interval. They localize the same target at the two interval endpoints; only the target specification changes.

Text

Input



Video frames

Prompt

You are given a sequence of frames uniformly sampled across a 52-second video. Each video frame is preceded by its timestamp in seconds. Complete the task based on the visual content of these video frames.

Question: How many points does Team USA player No. 7 score in total in the video?

A. 0 points
B. 2 points
C. 3 points
D. 4 points
E. 1 point

You must choose the most appropriate answer from the five options (A, B, C, D, or E). If uncertain, make your best choice based on the provided inputs. Output only the option letter.

Output

Ground Truth: `B`

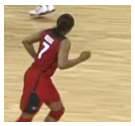
Text+Image

Input



Video frames

+



Reference image

Prompt

You are given a sequence of frames uniformly sampled across a 52-second video, followed by a reference image. Each video frame is preceded by its timestamp in seconds. Complete the task based on the visual content of the video frames and the reference image.

Question: How many points does the player in the reference image score in total in the video?

A. 0 points
B. 2 points
C. 3 points
D. 4 points
E. 1 point

You must choose the most appropriate answer from the five options (A, B, C, D, or E). If uncertain, make your best choice based on the provided inputs. Output only the option letter.

Output

Ground Truth: `B`

Figure E3: **Question Answering prompts for OmniCompass-VL.** Both settings receive the same uniformly sampled, timestamped video frames and retain the same answer choices and ground truth. Only the target specification changes.

Text

Input



Video frames with audio

Prompt

You are given a sequence of frames uniformly sampled across a 270-second video. Each frame is preceded by its timestamp in seconds. The audio track of the video is provided after the sampled frames. Complete the task based on the visual content of the sampled frames and the audio track.

Question: When in the video does the female news anchor with long brown curly hair, a black-and-white houndstooth jacket, and a black turtleneck report that France is seeking to redefine its role in Africa?

Output format: Answer with only this JSON object, where `start_time` and `end_time` are measured in seconds from the beginning of the video:

```
{"start_time": <number>,
"end_time": <number>}
```

Output

Ground Truth: `{"start_time": 89.0, "end_time": 92.0}`

Text+Image

Input



Video frames with audio + Reference image

Prompt

You are given a sequence of frames uniformly sampled across a 270-second video. Each frame is preceded by its timestamp in seconds. The audio track of the video is provided after the sampled frames, followed by a reference image. Complete the task based on the visual content of the sampled frames, the audio track, and the reference image.

Question: When in the video does the news anchor in the reference image report that France is seeking to redefine its role in Africa?

Output format: Answer with only this JSON object, where `start_time` and `end_time` are measured in seconds from the beginning of the video:

```
{"start_time": <number>,
"end_time": <number>}
```

Output

Ground Truth: `{"start_time": 89.0, "end_time": 92.0}`

Figure E4: **Temporal Grounding prompts for OmniCompass-AV.** Both settings receive the same timestamped video frames and corresponding audio, and predict the same audio-visual event interval. Only the target specification changes.

Text

Input



Video frames with audio

Prompt

You are given a sequence of frames uniformly sampled across a 270-second video. Each frame is preceded by its timestamp in seconds. The audio track of the video is provided after the sampled frames. Complete the task based on the visual content of the sampled frames and the audio track.

Question: Where in the frame is the female news anchor with long brown curly hair, a black-and-white houndstooth jacket, and a black turtleneck when she reports that French President Emmanuel Macron has arrived in Nairobi, Kenya?

The relevant interval starts at 85.03 seconds and ends at 87.87 seconds. Locate the target described in the question at each of these two timestamps and report one bounding box for each frame.
Write each box in `[x1, y1, x2, y2]` order, corresponding to `[left, top, right, bottom]`. Normalize all coordinates from 0 to 1000, with the top-left of the frame as the origin.

Output format: Return only this JSON object:

```
{"start_time": 85.03,
"start_bbox": [<x1>, <y1>,
<x2>, <y2>],
"end_time": 87.87,
"end_bbox": [<x1>, <y1>, <x2>,
<y2>]}
```

Output

Ground Truth:
```
{"start_time": 85.03,
"start_bbox": [278.96, 63.39, 749.11,
755.36],
"end_time": 87.87,
"end_bbox": [276.17, 68.35, 735.16,
760.32]}
```

Text+Image

Input



Video frames with audio + Reference image

Prompt

You are given a sequence of frames uniformly sampled across a 270-second video. Each frame is preceded by its timestamp in seconds. The audio track of the video is provided after the sampled frames, followed by a reference image. Complete the task based on the visual content of the sampled frames, the audio track, and the reference image.

Question: Where in the frame is the news anchor in the reference image when she reports that French President Emmanuel Macron has arrived in Nairobi, Kenya?

The relevant interval starts at 85.03 seconds and ends at 87.87 seconds. Locate the target described in the question at each of these two timestamps and report one bounding box for each frame.
Write each box in `[x1, y1, x2, y2]` order, corresponding to `[left, top, right, bottom]`. Normalize all coordinates from 0 to 1000, with the top-left of the frame as the origin.

Output format: Return only this JSON object:

```
{"start_time": 85.03,
"start_bbox": [<x1>, <y1>,
<x2>, <y2>],
"end_time": 87.87,
"end_bbox": [<x1>, <y1>, <x2>,
<y2>]}
```

Output

Ground Truth:
```
{"start_time": 85.03,
"start_bbox": [278.96, 63.39, 749.11,
755.36],
"end_time": 87.87,
"end_bbox": [276.17, 68.35, 735.16,
760.32]}
```

Figure E5: **Spatial Grounding prompts for OmniCompass-AV.** Both settings receive the same timestamped video frames, corresponding audio, and temporal interval. They localize the same target at the two interval endpoints; only the target specification changes.

Text

Input



Video frames with audio

Prompt

You are given a sequence of frames uniformly sampled across a 270-second video. Each frame is preceded by its timestamp in seconds. The audio track of the video is provided after the sampled frames. Complete the task based on the visual content of the sampled frames and the audio track.

Question: What happens at the end of the preceding report, just before the female news anchor with curly brown hair and a black-and-white patterned jacket reports that thousands of fans have gathered in Vienna?

A. Passengers disembark at a quay
B. A white-haired woman is embraced at an awards ceremony
C. Flag-waving fans crowd Vienna's streets
D. Two studio presenters discuss a political crisis
E. An evacuation plane waits on the apron

Choose the most appropriate answer from the five options (A, B, C, D, or E). If uncertain, make your best choice based on the provided inputs. Output only the option letter.

Output

Ground Truth: `B`

Text+Image

Input



Video frames with audio



Reference image

Prompt

You are given a sequence of frames uniformly sampled across a 270-second video. Each frame is preceded by its timestamp in seconds. The audio track of the video is provided after the sampled frames, followed by a reference image. Complete the task based on the visual content of the sampled frames, the audio track, and the reference image.

Question: What happens at the end of the preceding report, just before the news anchor in the reference image reports that thousands of fans have gathered in Vienna?

A. Passengers disembark at a quay
B. A white-haired woman is embraced at an awards ceremony
C. Flag-waving fans crowd Vienna's streets
D. Two studio presenters discuss a political crisis
E. An evacuation plane waits on the apron

Choose the most appropriate answer from the five options (A, B, C, D, or E). If uncertain, make your best choice based on the provided inputs. Output only the option letter.

Output

Ground Truth: `B`

Figure E6: **Question Answering prompts for OmniCompass-AV.** Both settings receive the same timestamped video frames, corresponding audio, and answer candidates. They share the same correct answer; only the target specification changes.