UNICBench: UNIfied Counting Benchmark for MLLM

Chenggang Rong[1*] Tao Han[2*] Zhiyuan Zhao[3*] Yaowu Fan[4]
Jia Wan[5] Song Guo[2] Yuan Yuan[1] Junyu Gao[1,†]

[1]Northwestern Polytechnical University [2]Hong Kong University of Science and Technology
[3]Institute of Artificial Intelligence (TeleAI), China Telecom [4]Sun Yat-sen University [5]Harbin Institute of Technology, Shenzhen

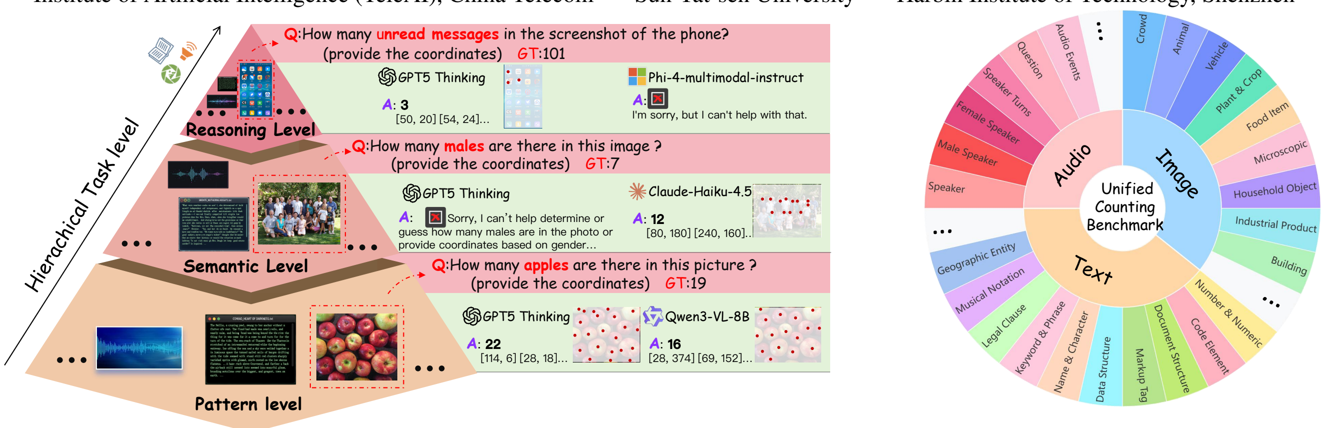

Figure 1. **Illustration of the benchmark's task taxonomy and dataset coverage**. The left pyramid groups counting problems by hierachical task level with representative Q/A examples. The right donut chart shows modality coverage and the diverse label categories in each modality, highlighting the benchmark's broad cross-modal and semantic coverage for unified counting evaluation.

Abstract

Counting is a core capability for multimodal large language models (MLLMs), yet there is no unified counting dataset to rigorously evaluate this ability across image, text, and audio. We present UNICBench, a unified multimodal, multi-level counting benchmark and evaluation toolkit with accurate ground truth, deterministic numeric parsing, and stratified reporting. The corpus comprises 5,300 images (5,508 QA), 872 documents (5,888 QA), and 2,069 audio clips (2,905 QA), annotated with a three-level capability taxonomy and difficulty tags. Under a standardized protocol with fixed splits/prompts/seeds and modality-specific matching rules, we evaluate 45 state-of-the-art MLLMs across modalities. Results show strong performance on some basic counting tasks but significant gaps on reasoning and the hardest partitions, highlighting long-tail errors and substantial headroom for improving general counting. UNICBench offers a rigorous and comparable basis for measurement and a public toolkit to accelerate progress.

1. Introduction

Counting is a core cognitive faculty and a key capability for multimodal large language models (MLLMs). It underlies number sense in humans and animals [12, 14]. As MLLMs progress toward more general, human-like intelligence [2, 7, 24, 54], reliable and interpretable counting across modalities becomes a concrete behavioral probe. Evaluating counting is both a practical test for applications (e.g., intelligent retail [50], security/surveillance [13, 19, 22, 38], scientometrics, audio analytics [4]) and a measure of how MLLMs approximate human numerical cognition.

Despite substantial progress of MLLMs on diverse multimodal benchmarks [62], including general-level visual question answering & reasoning (e.g., EvalQABench [66], SEED-Bench [30]), NoCaps [3]), (DocVQA [35], MMbench [33], MMMU [61]), and scientific question answering & reasoning (ScienceQA [41], MATH-Vision [51], ChemBench [21], SciBench [55]), a comprehensive, modality-spanning assessment of general counting remains lacking. In this paper, we benchmark MLLMs on diverse counting tasks that estimate the number of events, entities, or structural elements across images (e.g., people, vehicles, object instances [17, 29, 36, 60]), text (e.g., words, named entities, citations, paragraphs [56]), and audio

*Equal contribution.
†Corresponding author: gjy3035@gmail.com

(e.g., alarms, bird calls, percussive beats [43]), highlighting three complementary facets: perceptual localization, cross-span/multimodal aggregation and de-duplication, and rule-guided reasoning.

To rigorously assess how well today's MLLMs perform on general counting, a benchmark must address four challenges: (i) *modality and task coverage gaps*—key settings lack ready-to-use public data (e.g., audio events, image–text aligned counting, long-document structural element counts), requiring curated collection and human verification under licensing and privacy constraints; (ii) *heterogeneous annotations and no MLLM-ready QA standard*—existing datasets mix points, boxes, density maps, timestamps, and text spans with ambiguous instance inclusion rules, necessitating canonical target definitions and conversion to a unified QA template; (iii) *inconsistent evaluation protocols*—splits, prompts, decoding settings, matching rules, and randomness vary across works, hindering comparability and reproducibility; and (iv) *limited model availability and evaluation cost*—open models with balanced modality coverage are scarce, while closed-source APIs impose monetary costs, rate limits, and long-context token overheads, complicating fair comparisons.

Addressing these gaps requires a benchmark that is comprehensive in coverage, standardized in format, and rigorous in evaluation. We introduce a UNIfied Counting Benchmark (UNICBench) with the following contributions:

- We introduce the first unified *multimodal, multi-level* general-counting benchmark for MLLMs, extending visual counting to text and audio and formalizing a three-level capability and three-level difficulty taxonomy.
- We release a rigorously curated cross-modal corpus with evidence-first ground truth and a canonical schema for predictions and adapters, including 5,508 QA over 5,300 image samples, 5,888 QA over 872 text samples, and 2,905 QA over 2,069 audio clips.
- We evaluate state-of-the-art MLLMs (image: 21 models; text: 22 models; audio: 13 models) using unified protocols, revealing that current MLLMs perform well on simple tasks but require substantial improvement on reasoning and challenging tasks. The systematic analyses, consolidated reporting, and an open evaluation toolkit help facilitate future development of MLLM.

2. Related Work

2.1. Existing Datasets and Benchmarks

Image counting. Image counting has traditionally focused on crowded scenes (e.g., ShanghaiTech [63], UCF_CC_50 [27], NWPU-Crowd [53],GCC [52]), vehicle/traffic counting (e.g., CARPK [25], PUCPR+ [25]), and general object counting datasets (e.g., FSC-147 [39]). These datasets typically use density maps, point annotations, or bounding boxes and emphasize high-density and heavy-occlusion scenarios. However, they lack unified annotation formats (points/boxes/density) and consistent difficulty stratification, and few are naturally expressed in a question–answer format suitable for MLLMs [16, 64].

Text counting. Textual counting tasks appear across information extraction, document understanding, and scientometrics: from simple word, keyword counts and sentence, paragraph statistics to deduplicated citation counts and chart, table element counts, which appear as subproblems in DocVQA [35] and ChartQA [34]. Existing document QA benchmarks provide rich structural reading tests, but most do not focus on the semantic deduplication or cross-segment aggregation required by robust counting; annotations are typically stored as spans or indices and are not directly aligned to MLLM QA outputs [35].

Audio counting. Audio datasets are dominated by sound event detection and classification collections (e.g., AudioSet [20]). Audio counting tasks are challenging due to temporal overlap, short event durations, and varying annotation granularity, which complicates unified cross-modal evaluation [20, 48].

MLLM benchmarks. Recent multimodal benchmarks (VQA series [5, 8, 15], DocVQA [35], MMBench [33], MMMU [61], SEED-Bench [30], etc.) cover visual understanding, text recognition, structured reasoning, and audio–visual alignment. These benchmarks typically measure QA accuracy or retrieval metrics and do not systematically treat counting as a distinct capability across image/-text/audio simultaneously; a gap remains for an integrated, modality-spanning counting evaluation [30, 33, 61].

2.2. Existing Counting Methods

Classical counting. In computer vision, counting approaches fall into density estimation (CSRNet [31], ChfL [37], STEERER [23], et. al.), detection-then-count (LSC-CNN [42], TopoCount [1], et. al.), point-supervision methods (P2PNet [45], CLTR [32], APGCC [9], et. al.), and segmentation/ clustering based techniques. Density regression models (e.g., CSRNet variants) excel in ultra-dense crowds but are less robust on unseen scenes; detection-based pipelines (e.g., Faster-RCNN [40] + post-processing) work well on separable instances but struggle with occlusion. In audio, analogous approaches include sound event detection and temporal segmentation; in text, methods rely on information extraction (NER, coreference resolution) and rule-based aggregation.

LLM-based counting. With large language and multimodal models (Qwen2.5-Omni[57], Qwen3-Omni[58], Qwen2-Audio [10], Phi-Omni-ST [26]), researchers have explored prompting MLLMs for general counting, using chain-of-thought to elicit intermediate evidence, and building hybrid pipelines that combine detectors with an LLM

Table 1. Representative counting examples organized by category (capability vs. difficulty), level, and modality. The leftmost column groups three rows under "Capability level" (L1–L3) and a compact summary row for "Difficulty level" (Easy/Medium/Hard).

Category	Level	Image	Text	Audio
Capability level	Pattern (L1)	Q: "How many **people** are visible?"	Q: "How many times does **'Figure'** appear?"	Q: "How many **drum hits** in this clip?"
	Semantic (L2)	Q: "How many **people** are wearing **red shirts**?"	Q: "How many **non-repeated citations**?"	Q: "How many **bird calls** of **species X**?"
	Reasoning (L3)	Q: "How many **folders** in the specified path shown in the screenshot have a **modification date** in the **year 2022**?"	Q: "Count **references** from **2010 $\sim$ 2020**, **excluding appendices**."	Q: "In this audio, how many **questions** are **asked** in total?"
Difficulty level	Easy: **1–10** Medium: **11–100** Hard: **>100**	low density, small occlusion moderate density, some occlusion high density, severe occlusion	short span, low repetition moderate repetition, some deduplication long documents, heavy deduplication	sparse, low overlap moderate overlap, brief bursts overlapping events, long audio

for deduplication and rule aggregation. These approaches are flexible in zero- or few-shot settings but suffer from unverifiable intermediate evidence, poorly calibrated numeric confidences, and persistent cross-modal alignment failures [6, 65].

In summary, prior work provides numerous benchmarks and method families, but lacks a cross-modal, MLLM-centric benchmark that standardizes QA outputs, evidence reporting, and evaluation protocols for counting. Our work addresses this gap by proposing a unified benchmark.

3. UNIfied Counting Benchmark

3.1. Benchmark Overview

Task taxonomy. We organize all counting tasks into capability levels and difficulty tags. Table 1 gives representative examples for each modality and capability level, illustrating typical question templates. To strictly demarcate these levels, we classify the query Q based on its required operation on the entity set E:

- **Pattern level (L1)** — perceptual counting: direct observation of instances/events suffices; no semantic filtering or rule application is required ($y = |E|$).
- **Semantic level (L2)** — attribute filtering and deduplication: Q applies atomic filters P on *intrinsic attributes* (e.g., color, type) or performs identity aggregation (counting unique entities), yielding $y = |\{e \in E \mid P(e)\}|$.
- **Reasoning level (L3)** — rule-driven and compositional counting: Q imposes *explicit rules* or *structural constraints*, such as arithmetic/logical weights ($y = g(|S_1|, \dots)$) or temporal/complex interactions ($y = |\{e \mid \text{TempStruct}(e)\}|$).

In addition, each sample is also tagged with a difficulty label (**Easy/Medium/Hard**). As shown in Table 1, difficulty is quantified by measurable attributes (e.g., object number) to enable stratified analysis.

Question-Answer guidelines. To ensure consistency across modalities and levels, QA follows these rules:

- **Level assignment**: Annotators assign L1–L3 strictly following the taxonomy defined above, ambiguous cases are adjudicated by lead annotators.
- **Difficulty quantification**: Compute objective measures (density, occlusion, overlap ratio, cross-segment repetition) and map to Easy/Medium/Hard thresholds; store these measures in the data card for each sample.
- **Evidence-first GT**: Every ground truth includes both gt_count and structured gt_evidence (points/spans/timestamps).
- **Question templates**: We use deterministic templates for L1 tasks (to minimize linguistic variance), and more free-form formulations for L2/L3 that specify filtering rules or aggregation constraints explicitly.

3.2. Data Collection

Data Sources. The benchmark aggregates data from multiple sources. Image samples are drawn from established counting datasets (FSC147, NWPU-MOC[18], CARPK, JHU-CROWD++[44], UCF-QNRF[28], ShanghaiTech, IOCfish5K[46], Global Wheat Head Detection[11], Snapshot Serengeti[47]), with additional manual annotations for categories like screen panels, pens, birds, seagulls, books, biscuits, chairs, donuts tray, marbles, cups, mini blinds, potatoes, alcohol bottles, crows, green peas, and lipstick. Text data (5,888 questions across 872 samples) are entirely self-collected from diverse corpora including code repositories, legal documents, academic LaTeX files, and literary works. Audio samples leverage DESED [49] for environmental sounds and AliMeeting [59] for conversational speech.

Preprocessing Pipelines. For images, we perform deduplication, quality filtering, and annotation format unification, converting point annotations to instance coordinates while preserving resolution diversity (from 234×180 to 6736×4640 pixels). Text preprocessing includes deduplication, segmentation, and character-span annotation, with ancient and literary texts retaining original linguistic structures. Audio processing applies noise reduction, temporal segmentation, and precise event timestamp alignment, ensuring sub-second temporal accuracy.

Quality Control. We employ a multi-stage verification protocol: dual independent annotation with arbitration for disagreements, achieving 100% annotation consistency. Random sampling reviews ensure cross-modal annotation

Figure 2. a) Sample visualization in three modalities. b) Overview of dataset composition. Sample and question distributions across modalities and the capability/difficulty breakdown. c) Smoothed ground-truth count distributions for three modalities, which are skewed and long-tailed and thus motivate our stratified difficulty thresholds and evaluation protocol. d) Category word cloud based on question counts. e) f) g) Distribution of resolution/text length/ audio duration in three modalities, respectively.

coherence. Questions incorporate deduplication challenges (cross-sentence entity counting), difficulty stratification via the three-level taxonomy, and adversarial prompts to test reasoning robustness.

Ethics and Licensing. All ata adhere to original licenses and are anonymized to protect privacy.

3.3. Data Distribution

Figure 2 a) summarizes the corpus composition. Figure 2 b) presents the comprehensive dataset composition across modalities, sources, and difficulty levels. There are two distributional patterns: First, sample-level coverage and question-level coverage are not identical. While image samples constitute the largest share of raw items, the number of questions is more evenly distributed across image and text modalities. This gap reflects our collection strategy of producing multiple, diverse QA probes per text sample (e.g., several citation/paragraph questions per paper) while keeping image samples closer to one QA per scene.

Image counting track. The 5,300 image samples (5,508 questions) exhibit diverse scene complexity. Count ranges span from sparse scenarios (airplane: average 5.58, stadium: 3.8) to extremely dense scenes (crowd: 355.58, tree: 301.96, fish: 135.12). Resolution distribution varies from 234×180 to 6736×4640 pixels (mean: 804×660, aspect ratio: 1.28), accommodating both high-resolution remote sensing and standard digital photography. Category distribution emphasizes everyday objects (30+ categories from FSC147), aerial imagery (12 categories from NWPU-MOC), dense crowds (150 samples), and specialized domains (cells, wildlife). Difficulty stratification concentrates on Pattern-level tasks (85.4%), reflecting the predominance of direct visual counting, with Semantic (10.0%) and Reasoning (4.6%) levels addressing attribute filtering and spatial relationships.

Text counting track. The 872 text samples (5,888 questions) display extreme length variability (584–8,052,974 characters, median: 7,176). Categories distribute across code (100 samples), HTML (168), JSON (200), Musical Notation (143), LaTeX (66), News (60), Literary (60), CSV (15), Exam papers (15), Official documents (15), ancient texts (20), and law (10). Ancient and Literary categories preserve original language (classical Chinese and source language respectively), totaling 80 bilingual samples (9.2%). Count distributions exhibit high variance (mean: 120.40, median: 4.0, max: 71,176), with LaTeX and CSV categories contributing extreme values. Difficulty distribution uniquely favors Reasoning-level (43.7%) and Semantic-level (29.9%) tasks, demanding structural understanding and cross-reference deduplication.

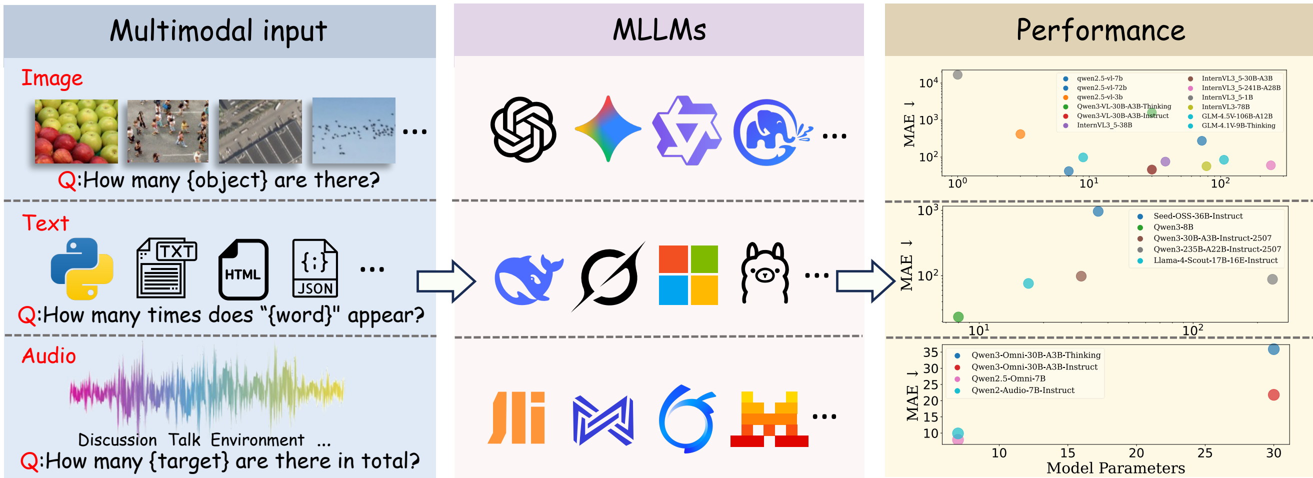

Figure 3. Overview of the **UNICBench** pipeline. We standardize multi-modal datasets and define a unified QA-evidence schema with difficulty and capability labels. The end-to-end framework enables assessment of MLLM counting across image, text, and audio.

Audio counting track. The audio track (2,069 samples, 2,905 QA) covers environmental (1,860 samples from DESED dataset [49]) and conversational speech (209 samples from AliMeeting [59]). Durations range from 3.0 to 2,497.9 seconds (median: 10.0). Environmental audio exhibits sparse event density (1.56 events/sample), while conversational audio shows dense segmentation (81.51 counts/sample). All audio maintains WAV format with varied sampling rates preserving source characteristics. Difficulty balances Pattern-level (64.0%) direct event counting with Semantic/Reasoning-level (36.0%) speaker analysis and turn-taking inference.

4. Evaluation

Figure 3 depicts an overview of the UNICBench. The evaluation is performed on three tracks with corresponding MLLMs. Following, we will introduce the evaluation protocols, results, and analysis of each track. *More details and settings are provided in the Supplementary.*

4.1. Metrics

We evaluate counting performance with diverse metrics. Let N be the number of evaluation examples, y_i the ground-truth count for example i, and $\hat{y}_i$ the model prediction parsed as a numeric value. If a model response is not parsable as a number (e.g., out-of-context, timeout, or error), we treat $\hat{y}_i$ as invalid for metrics that require numeric predictions; such cases are only counted in Success Rate.

Success rate (robustness). It measures the fraction of examples for which the model returns a parsable numeric prediction. Let $\mathbf{v}_i = 1$, if $\hat{y}_i$ is a valid numeric prediction, else $\mathbf{v}_i = 0$, Then

$$\text{SuccessRate} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{v}_i. \tag{1}$$

All subsequent statistical metrics are based on the successful parsing of results.

Mean Absolute Error and Mean Squared Error.

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i|, \quad \text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2. \tag{2}$$

Hit rate (tolerance based). For a relative tolerance τ (e.g., $\tau = 0.10$ for 10%), define the hit indicator

$$\mathbf{1}_i^{(\tau)} = \begin{cases} 1 & \text{if } \dfrac{|\hat{y}_i - y_i|}{\max(1, y_i)} \le \tau, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

The Hit Rate at tolerance τ is

$$\text{HitRate@}(1-\tau) = \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}_i^{(\tau)}. \tag{4}$$

We report HitRate@100% (Estimated number is equal to annotation), HitRate@90% and HitRate@80%.

4.2. General Setting

System Prompt. We use a consistent system prompt across all models to ensure a uniform understanding:

```
You are a counting assistant. You MUST respond with
ONLY a number. Never refuse to answer. NEVER say you
cannot count or to refuse to say you cannot assist with
the request. Always give your best numerical estimate.
Respond with just the number, nothing else.
```

Answer Extraction. During answer extraction we required that the model return a single numeric value. To accommodate models that produce internal deliberation tokens or proprietary wrappers, we accept (and strip) explicit thinking tags like <think> and final-answer delimiters like <answer> (and also common vendor wrappers

Table 2. **Benchmark results on the Image-modality counting track**. Metrics are: SuccessRate (%), Hit rates (@100%/@90%@80%), MAE/MSE for Overall, per-difficulty, and per-capability. MSE values are shown in scientific notation with one decimal.

Model	Success (%)	Hit rate (%)			Overall		Easy		Medium		Hard		Pattern		Reasoning		Semantic	
		@100% ↑	@90% ↑	@80% ↑	MAE ↓	MSE ↓	MAE ↓	MSE ↓	MAE ↓	MSE ↓	MAE ↓	MSE ↓	MAE ↓	MSE ↓	MAE ↓	MSE ↓	MAE ↓	MSE ↓
Claude-Sonnet-4-20250514	100.0	14.9	27.8	42.4	78.1	1.1e6	5.4	6.1e3	16.8	4.5e3	444.6	7.0e6	68.8	1.2e6	4.4	5.7e1	191.6	4.1e5
Gemini-2.5-Flash	100.0	15.6	25.1	38.0	140.5	1.0e6	12.0	5.7e3	55.8	2.9e4	694.2	6.5e6	131.4	1.0e6	6.7	2.1e3	280.8	1.1e6
Gemini-2.5-Pro-Thinking	100.0	20.3	37.5	51.3	90.0	8.1e5	4.3	5.8e2	22.1	2.9e4	504.9	5.3e6	71.1	6.5e5	4.6	8.4e1	291.1	2.5e6
GPT-4o-mini	100.0	12.5	20.7	33.0	73.3	1.5e6	2.3	**3.8e1**	15.0	5.0e2	424.6	9.7e6	72.7	1.7e6	5.3	7.4e1	109.8	5.7e4
GPT-5	100.0	16.8	32.3	48.3	54.1	2.0e6	2.5	2.1e2	11.0	**3.1e2**	312.4	1.4e7	55.1	2.4e6	5.9	1.0e2	**67.4**	3.3e4
GPT-o3	100.0	18.5	33.6	49.7	49.0	5.0e5	2.8	6.6e2	11.2	4.5e2	277.1	3.3e6	44.3	5.4e5	4.4	8.5e1	109.2	3.3e5
GPT-4o	100.0	16.6	31.6	48.1	43.2	7.5e5	2.4	1.3e2	11.4	4.7e2	238.4	5.0e6	41.7	8.7e5	5.4	1.1e2	73.7	4.7e4
o4-mini	100.0	17.1	32.3	48.7	42.9	9.4e5	2.2	8.1e1	10.7	4.6e2	239.1	6.2e6	39.1	1.1e6	4.1	6.9e1	92.5	2.0e5
GPT-5-mini	100.0	17.2	32.8	50.5	**29.8**	9.6e4	2.1	6.5e1	**10.0**	3.4e2	**155.0**	6.4e5	**25.4**	1.1e5	5.3	1.1e2	78.9	2.7e4
GLM-4.1V-9B-Thinking	87.3	15.7	27.5	43.2	97.9	1.6e6	3.0	4.4e2	13.9	1.1e3	542.2	9.9e6	90.0	1.7e6	**3.1**	**2.5e1**	207.5	1.7e6
GLM-4.5V-106B-A12B	100.0	16.0	28.4	44.1	84.5	3.5e6	2.7	2.4e2	12.4	1.0e3	509.8	2.3e7	89.5	4.1e6	4.3	6.5e1	78.8	**2.3e4**
InternVL3_5-1B	100.0	9.3	15.1	26.1	16686.6	1.5e12	4.5	2.2e3	19.1	1.6e3	111186.9	1.0e13	19506.0	1.8e12	6.7	1.0e2	346.6	1.5e7
InternVL3_5-38B	100.0	16.6	27.8	42.9	75.4	1.3e6	2.3	7.2e1	15.9	6.7e2	435.5	8.5e6	73.2	1.5e6	6.8	1.1e2	126.3	3.3e5
InternVL3_5-241B-A28B	100.0	**21.9**	**40.5**	**57.1**	59.9	2.9e6	1.9	5.4e1	10.0	8.0e2	356.4	2.0e7	61.1	3.4e6	6.1	9.3e1	74.9	3.0e4
InternVL3-78B	99.6	17.8	28.8	47.9	56.6	2.1e6	**1.7**	4.0e1	12.5	5.0e2	326.0	1.4e7	54.3	2.4e6	4.9	7.0e1	99.7	1.6e5
InternVL3_5-30B-A3B	100.0	19.1	32.4	44.6	46.1	**4.5e4**	1.9	5.6e1	17.5	8.1e2	234.2	**3.0e5**	41.0	4.1e4	6.4	1.0e2	108.3	1.0e5
Qwen3-VL-30B-A3B-Thinking	92.7	18.3	30.8	45.5	1583.1	1.1e9	2.3	1.3e2	20.2	1.2e4	10293.5	7.5e9	355.2	6.7e7	3.5	4.9e1	13354.7	1.1e10
Qwen2.5-VL-3B	100.0	8.9	15.9	29.1	417.7	3.6e8	3.5	9.9e1	21.0	2.4e3	2696.1	2.4e9	36.7	**2.2e4**	5.4	1.2e2	3860.4	3.6e9
Qwen2.5-VL-72B	100.0	17.4	32.8	48.8	274.8	1.9e8	2.4	9.6e1	12.3	9.8e2	1779.1	1.2e9	92.7	4.4e6	4.5	7.5e1	1953.2	1.8e9
Qwen3-VL-30B-A3B-Instruct	100.0	17.3	27.1	39.5	45.7	7.3e4	1.9	3.8e1	13.8	4.3e2	246.7	4.8e5	37.3	7.1e4	5.4	8.3e1	136.1	1.2e5
Qwen2.5-VL-7B	99.9	16.6	28.6	43.2	41.8	1.1e5	2.0	5.9e1	13.6	4.6e2	222.3	7.5e5	37.6	1.3e5	4.5	7.0e1	95.0	5.4e4

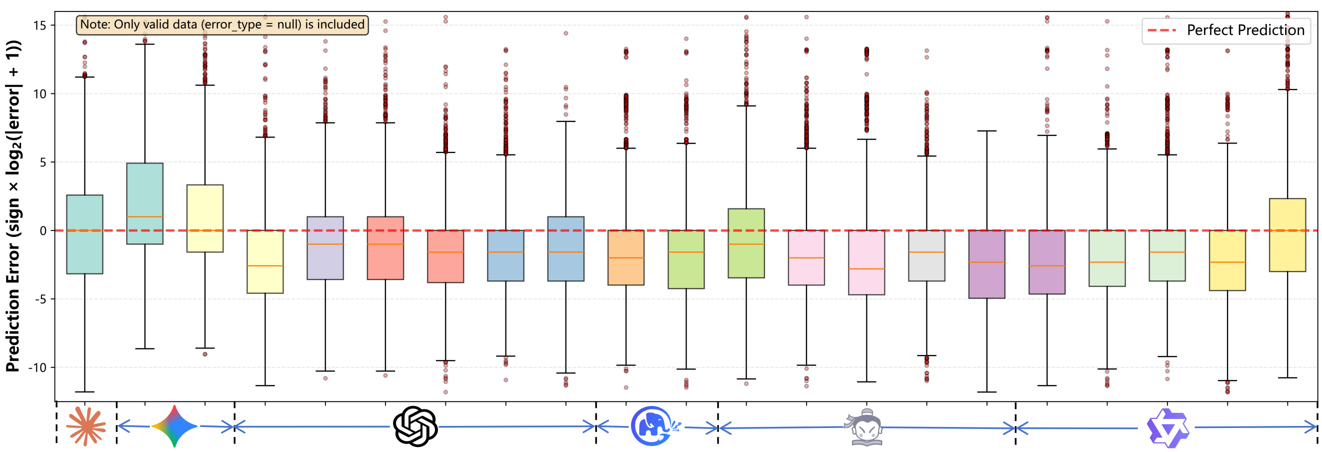

Figure 4. **Distribution of prediction error on image modality**. Whiskers and outliers indicate extreme failures—long whiskers or many outliers show a model makes severe errors on some samples (e.g., rare classes, label noise, or collapse cases). Such frequent extreme errors can substantially increase MAE/MSE even when the median error looks small.The model ordering is consistent with that in Table 2.

such as GLM's <begin_of_box>...</end_of_box>). Our parser priority is: (1) extract a numeric token inside <answer> if present; (2) otherwise accept a standalone numeric token at the end of the response; (3) otherwise fall back to the first parseable numeric token in the output.

4.3. Image Counting Track

Setting. For the image-counting evaluation we enforced a unified runtime configuration across models to ensure comparability. Inference parameters were standardized to: {max_tokens:4096, temperature:0.0, timeout_seconds: 120}. To balance responsiveness and computational cost, certain models were run with reduced reasoning intensity (e.g., GPT-5 and GPT-5-mini with {reasoning_effort: minimal, text_verbosity: low}; GPT-o3 set reasoning_effort to low). A few models that do not expose temperature were executed at their default temperature (notably GPT-5, GPT-5-mini, GPT-4o-mini and GPT-o3 at temperature = 1.0).

Results. Table 2 reports per-model performance on the image-counting track. Success rates are high (many models reach 100%), indicating that producing numeric outputs per prompt is not the primary bottleneck. Hit rates rise with wider tolerance (100%→90%→80%), yet only three models exceed a 50% hit rate at the 20% error threshold (InternVL3_5-241B-A28B, Gemini-2.5-Pro-Thinking, and GPT-5-mini), indicating that current MLLMs still have substantial room to improve absolute counting accuracy. Error trends across difficulty levels are monotonic: MAE and MSE increase from Easy to Medium to Hard. Top-tier closed-source models (e.g., GPT-5-mini, GPT-o4-mini) show less relative error inflation on Hard, indicating greater robustness, while several open-source models (e.g., Qwen2.5-VL-7B and InternVL3_5-30B-A3B) remain competitive in dense scenes. By capability, Reasoning tasks yield the lowest numeric errors, largely because they often have small true counts, whereas Pattern and Semantic

Table 3. **Benchmark results on the Text-modality counting track**. Metrics are: SuccessRate (%), Hit rates (@100%/@90%@80%), MAE/MSE for Overall, per-difficulty, and per-capability. MSE values are shown in scientific notation with one decimal.

Model	Success (%)	Hit rate (%)			Overall		Easy		Medium		Hard		Pattern		Reasoning		Semantic	
		@100% ↑	@90% ↑	@80% ↑	MAE ↓	MSE ↓	MAE ↓	MSE ↓	MAE ↓	MSE ↓	MAE ↓	MSE ↓	MAE ↓	MSE ↓	MAE ↓	MSE ↓	MAE ↓	MSE ↓
Gemini-2.5-Flash-Nothinking	99.0	37.3	43.7	50.5	67.0	3.4e5	7.4	2.6e4	24.1	7.1e4	746.3	4.1e6	33.3	1.7e4	120.1	7.2e5	19.8	7.3e4
Gemini-2.5-Pro-Thinking	99.2	**63.3**	**72.8**	**76.1**	30.5	9.0e4	3.6	4.5e2	11.3	1.0e4	**337.4**	**1.2e6**	10.5	3.2e3	56.4	2.0e5	10.6	1.1e4
GPT-5-mini	97.6	35.3	40.6	48.6	1665.2	1.4e10	2347.8	2.1e10	17.9	1.1e4	1321.4	2.9e7	20.8	4.4e3	3804.8	3.2e10	8.1	2.2e3
GPT-o3	96.8	61.1	68.2	72.0	112.9	2.3e6	3.5	4.7e2	16.4	2.5e4	1510.8	3.2e7	15.5	1.0e4	245.6	5.2e6	5.6	1.5e3
GPT-5	97.8	36.2	39.7	47.2	106.9	1.2e6	6.0	5.7e2	16.6	2.5e3	1393.6	1.7e7	23.6	4.2e3	225.8	2.7e6	7.0	1.8e3
GPT-4o-mini	97.2	35.5	37.4	41.6	80.8	5.3e5	2.9	7.6e1	24.4	3.0e4	1057.7	7.7e6	29.3	5.8e3	162.1	1.2e6	8.2	3.1e3
GPT-4o	97.1	39.8	41.8	46.2	63.0	3.6e5	4.6	3.8e2	18.4	6.6e3	811.6	5.3e6	26.1	2.9e3	122.1	8.3e5	9.8	7.3e3
o4-mini	97.6	59.6	67.1	70.2	47.1	2.2e5	4.3	3.6e2	9.8	1.1e3	594.0	3.2e6	15.1	1.9e3	95.5	5.1e5	5.0	1.6e3
Grok-4-Fast-Non-Reasoning	94.5	37.6	42.2	47.0	67.0	1.1e6	5.7	5.5e2	16.9	5.8e3	895.3	1.7e7	90.2	3.3e6	90.6	5.1e5	14.5	1.9e4
Claude-Sonnet-4-20250514	97.6	37.7	44.3	50.4	60.0	4.6e5	3.7	3.5e2	18.7	1.2e4	753.4	6.5e6	19.3	2.6e3	120.9	1.0e6	6.9	1.2e3
DeepSeek-V3.1-Nothinking	97.1	40.1	46.0	51.2	75.3	4.2e5	3.2	4.0e2	20.9	1.5e4	998.2	6.2e6	30.0	7.9e3	149.7	9.7e5	7.0	1.6e3
Deepseek-V3.2-exp	97.5	34.7	38.1	43.9	71.2	7.0e5	5.1	6.5e2	27.3	1.8e4	876.5	1.0e7	35.3	1.5e4	135.4	1.6e6	9.1	2.4e3
DeepSeek-V3.1	94.2	47.3	51.7	55.8	46.9	1.5e5	2.8	1.3e2	16.7	1.4e3	653.0	2.5e6	29.2	7.0e3	86.5	3.5e5	7.0	3.4e3
DeepSeek-R1-0528	94.7	59.7	67.8	71.3	34.6	8.5e4	4.2	3.0e2	**8.6**	6.9e2	469.0	1.3e6	15.4	1.8e3	68.0	2.0e5	4.4	9.0e2
GLM-4.6	100.0	32.0	34.4	39.4	151.3	2.8e6	6.0	5.5e2	30.6	3.1e4	1915.3	3.8e7	108.0	3.4e6	272.3	4.4e6	13.5	1.4e4
Llama-4-Scout-17B-16E-Instruct	98.2	31.6	33.7	37.6	76.2	6.2e5	4.5	1.8e2	17.4	9.8e2	932.4	8.5e6	37.2	2.3e4	140.2	1.4e6	18.8	2.8e4
Phi-4-Multimodal-Instruct	88.4	21.3	22.5	26.6	82.7	3.2e5	10.0	2.3e3	23.2	1.8e3	1132.4	5.2e6	115.1	3.1e5	115.0	5.5e5	7.7	8.1e2
Phi-4	69.7	41.9	42.6	45.3	**9.6**	**1.1e4**	**1.5**	**1.7e1**	15.8	4.1e2	551.3	1.4e6	**8.9**	**4.5e2**	**13.5**	**2.3e4**	5.5	2.6e3
Qwen3-30B-A3B-Instruct-2507	94.8	28.1	29.8	33.1	98.4	9.3e5	3.7	1.2e2	31.4	1.6e4	1207.8	1.3e7	39.7	7.6e3	194.1	2.1e6	8.1	1.9e3
Qwen3-235B-A22B-Instruct-2507	97.4	36.2	38.4	43.6	88.2	1.4e6	3.1	1.4e2	23.4	2.0e4	1161.0	2.0e7	29.8	5.6e3	180.4	3.1e6	6.7	1.6e3
Qwen3-8B	68.8	53.3	58.6	64.9	23.4	7.3e4	2.5	9.3e1	10.4	**3.1e2**	675.1	2.6e6	9.1	9.9e2	46.9	1.7e5	**3.6**	**3.8e2**
Seed-OSS-36B-Instruct	95.6	35.3	37.8	42.3	966.8	1.2e8	9.7	2.1e3	63.5	2.8e5	13616.3	1.8e9	115.8	2.9e5	2193.7	2.9e8	13.3	2.1e4

tasks remain challenging under dense packing or occlusion. High-density categories (e.g., crowd, tree, fish) and large scale variation primarily drive extreme errors. Increasing image resolution helps these cases but yields limited gains in ultra-dense settings due to representation bottlenecks.

Analysis. Figure 4 presents each model's error distribution with boxplots. Frontier closed-source and large open models exhibit narrower interquartile ranges (IQRs) and fewer extreme outliers, whereas lightweight models display wider dispersion and heavier tails. Signed-error medians are often negative, and positive tails indicate rare large overestimates that inflate MAE and MSE. We attribute these patterns to three interacting factors:

- *Supervision gap.* Models are optimized for instruction following and description rather than dense instance-level counting, and thus lack per-instance supervision (e.g., points or detections) needed to resolve tightly clustered or occluded instances.
- *Representation limits.* Patch-based, downsampled visual tokens compress small or densely packed objects, undermining one-to-one attention and increasing misses and duplicates.
- *Calibration and distribution shift.* Decoding preferences favor rounding under uncertainty and occasionally overconfident predictions; combined with pretraining corpora that under-represent high-density or occluded domains, these factors amplify long-tailed errors.

4.4. Text Counting Track

Setting. Apart from the additional changes mentioned, other model settings were kept consistent with image counting track. A few models used the officially recommended temperature parameters:{GLM-4.6:1.0, Phi-4:1.0, and Phi-4-Multimodal-Instruct:0.75}. We also evaluated GLM-4.6 in a "non-thinking" (no internal reasoning) mode to assess

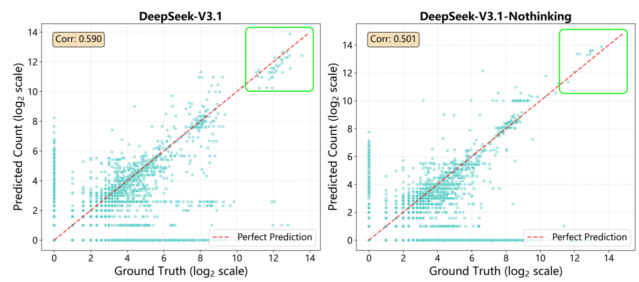

Figure 5. **Prediction vs. ground truth scatter.** Each point is one sample. The upper-right green box highlights a small set of extreme high-count samples where the thinking mode defeated non-thinking mode, which drives most of the overall MAE gap.

the impact of explicit chain-of-thought behavior.

Results. Table 3 summarizes performance on the text counting track. Extraction success is high overall, whereas numeric accuracy varies widely. Across our runs, Phi-4, Qwen3-8B, and Gemini-2.5-Pro-Thinking achieve the lowest aggregate MAE/MSE; several GPT and Qwen variants are competitive, but a few models show anomalous biases that inflate errors. As with images, errors increase monotonically with difficulty (Easy→Medium→Hard). Closed-source models maintain lower MAE/MSE on Hard, indicating greater robustness, though some large open models (e.g., DeepSeek-R1-0528) also perform strongly. By capability, Semantic tasks yield the smallest errors, indicating that modern MLLMs handle semantic filtering and aggregation well. Pattern-level tasks requiring strict template or format matching are more error-prone, and Reasoning remains the principal bottleneck and the largest source of inter-model variance.

Analysis. Figure 5 compares DeepSeek-V3.1 and DeepSeek-V3.1-Nothinking in text counting. The models perform similarly on most samples, butextreme high-count

Table 4. **Benchmark results on the audio-modality counting track**. Metrics are: SuccessRate (%), Hit rates (@100%/@90%@80%), MAE/MSE for Overall, per-difficulty, and per-capability. MSE values are shown in scientific notation with one decimal. A “-” indicates that no valid values were obtained for this statistical dimension.

Model	Success (%)	Hit rate (%)			Overall		Easy		Medium		Hard		Pattern		Reasoning		Semantic	
		@100% ↑	@90% ↑	@80% ↑	MAE ↓	MSE ↓	MAE ↓	MSE ↓	MAE ↓	MSE ↓	MAE ↓	MSE ↓	MAE ↓	MSE ↓	MAE ↓	MSE ↓	MAE ↓	MSE ↓
Gemini-2.5-Pro-Thinking	64.0	31.0	31.0	31.0	1.1	2.3e0	1.1	2.3e0	-	-	-	-	1.1	2.3e0	-	-	-	-
Gemini-2.5-Flash-Nothinking	64.0	**48.0**	**48.0**	**48.0**	0.7	1.0e0	0.7	1.0e0	-	-	-	-	0.7	1.0e0	-	-	-	-
GPT-4o-mini-Audio-Preview	64.0	11.3	11.3	11.3	1.7	4.1e0	1.7	4.1e0	-	-	-	-	1.7	4.1e0	-	-	-	-
GPT-4o-Audio-Preview	63.9	32.4	32.4	32.4	0.9	1.3e0	0.9	1.3e0	-	-	-	-	0.9	1.3e0	-	-	-	-
GPT-Audio	64.0	33.9	33.9	33.9	0.9	1.5e0	0.9	1.5e0	-	-	-	-	0.9	1.5e0	-	-	-	-
GPT-Audio-mini	63.0	38.3	38.3	38.3	**0.6**	**6.8e-1**	**0.6**	**6.8e-1**	-	-	-	-	**0.6**	**6.8e-1**	-	-	-	-
Voxtral-mini	99.8	36.1	36.1	36.1	29.5	7.2e3	1.8	5.1e1	74.8	5.9e3	222.2	6.0e4	0.9	3.9e0	194.3	5.0e4	4.3	1.9e2
Voxtral-small	73.4	29.5	29.5	29.5	1.5	4.9e1	1.1	2.1e0	74.0	5.5e3	**122.6**	**1.5e4**	1.1	1.9e0	**101.0**	**1.1e4**	**1.1**	3.7e0
Phi-4-Multimodal-Instruct	63.7	2.0	2.0	2.0	3.2	1.3e1	3.2	1.3e1	-	-	-	-	3.2	1.3e1	-	-	-	-
Qwen2-Audio-7B-Instruct	99.7	9.9	9.9	9.9	30.1	7.3e3	2.4	8.3e0	**70.3**	**5.2e3**	223.4	6.1e4	2.8	1.0e1	194.5	5.0e4	1.3	**2.4e0**
Qwen3-Omni-30B-A3B-Instruct	100.0	21.8	21.8	21.8	29.6	7.2e3	1.8	9.1e0	74.0	5.8e3	223.2	6.0e4	1.0	2.0e0	195.0	5.0e4	4.0	3.0e1
Qwen2.5-Omni-7B	100.0	7.9	7.9	7.9	29.2	7.1e3	1.8	4.6e0	70.9	5.3e3	220.7	6.0e4	1.8	4.0e0	192.4	4.9e4	1.8	6.4e0
Qwen3-Omni-30B-A3B-Thinking	73.5	35.6	35.7	35.9	14.7	8.6e3	6.9	4.5e3	79.5	4.7e4	310.0	1.6e5	0.9	1.8e0	249.9	1.3e5	59.5	4.4e4

cases drive the MAE gap; in the right panel's top-right green box, the thinking variant shows smaller errors. Consequently, its overall MAE is 46.9 versus 75.3 for the non-thinking model (Table 3), indicating that improvements on a few long-tail, large-count samples can shift aggregate error despite similar median performance.

4.5. Audio Counting Track

Setting. This track follows the same base runtime protocol as previous tracks. All models were evaluated using the common configuration {max_tokens:4096, temperature:0.0, timeout_seconds:120}.

Results. Table 4 summarizes the audio-track results; only Qwen2.5-Omni-7B and Qwen3-Omni-30B-A3B-Instruct achieve perfect extraction success. Several models, including GPT-family variants and Phi-4-Multimodal-Instruct, record success rates near 60%, yet their numeric outputs often attain competitive MAE/MSE. Audio-specialized GPT variants (e.g., GPT-Audio-mini, GPT-Audio) and Gemini models demonstrate strong numeric precision, consistent with audio-focused pretraining. Within the Voxtral family, Voxtral-mini outperforms Voxtral-small; Qwen Audio-enabled models exhibit high success but are prone to systematic overestimation, increasing aggregate error. Difficulty-stratified metrics are monotonic (Easy→Medium→Hard), with the largest inter-model differences on Hard examples; by capability, Pattern and Semantic tasks are relatively easy, whereas Reasoning tasks yield the greatest separation across models.

Analysis. Figure 6 examines the relationship between refusal behavior and MAE. Models such as Qwen2.5-Omni-7B and Voxtral-mini achieve near-perfect extraction success rates (see Table 4, 100.0% and 99.8%, respectively), and therefore they produce numeric answers even for the most challenging samples; consequently, their overall MAE lies in the tens (29.2 and 29.5), primarily driven by a small number of extreme, high-count cases. In contrast, several models with lower success rates (e.g., certain GPT-Audio variants) exhibit MAE values below 1 because many difficult items result in refusals, input failures, or non-numeric

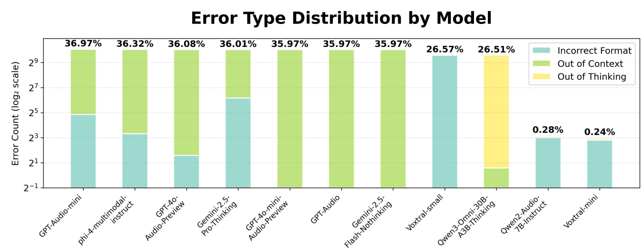

Figure 6. **Error-type analysis on the audio modality**. Models that refuse or fail to produce numeric outputs more often avoid counting those errors in MAE (but are penalized in SuccessRate).

outputs, a pattern also reflected in the SuccessRate metric. Notably, Qwen2.5-Omni-7B's combination of a high SuccessRate and relatively low MAE demonstrates robust and accurate counting under our unified evaluation protocol.

For more evaluation results, details and analysis, please refer to the Evaluation Detail section in the Supplementary material.

5. Conclusion

UNICBench is a unified multimodal counting benchmark spanning image, text, and audio with a three-tier taxonomy (Pattern, Semantic, Reasoning), stratified difficulty thresholds, and evidence-first ground truth; our experiments show that modern multimodal LLMs handle many Pattern and Semantic tasks well but still struggle on Reasoning and hardest partitions, with modality-specialized models often offering higher numeric precision; based on these findings we recommend evidence-first outputs, hybrid pipelines that combine detectors and VLLMs, and difficulty-stratified reporting, while noting limitations from dataset bias and evolving tooling; future work will focus on improved cross-modal alignment, tighter detector integration, and automated calibration for robust numeric reasoning.

Acknowledgments. This work was supported in part by grants from the National Natural Science Foundation of China (62576284 & 62306241).

References

[1] Shahira Abousamra, Minh Hoai, Dimitris Samaras, and Chao Chen. Localization in the crowd with topological constraints. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(2):872–881, 2021. 2

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1

[3] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8948–8957, 2019. 1

[4] Amirreza Ahmadnejad, Ahmad Mahmmodian Darviishani, Mohmmad Mehrdad Asadi, Sajjad Saffariyeh, Pedram Yousef, and Emad Fatemizadeh. Tacnet: Temporal audio source counting network. *arXiv preprint arXiv:2311.02369*, 2023. 1

[5] A. Alfarano, L. Venturoli, and D. Negueruela del Castillo. Vqart-bench: A semantically rich vqa benchmark for art and cultural heritage. *arXiv preprint arXiv:2510.12750*, 2025. 2

[6] Iñigo Alonso, Gorka Azkune, Ander Salaberria, Jeremy Barnes, and Oier Lopez de Lacalle. Vision-language models struggle to align entities across modalities. *arXiv preprint arXiv:2503.03854*, 2025. 3

[7] 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*, 2025. 1

[8] James Burgess, Jeffrey J Nirschl, Laura Bravo-Sánchez, Alejandro Lozano, Sanket Rajan Gupte, Jesus G. Galaz-Montoya, Yuhui Zhang, Yuchang Su, Disha Bhowmik, Zachary Coman, Sarina M. Hasan, Alexandra Johannesson, William D. Leineweber, Malvika G Nair, Ridhi Yarlagadda, Connor Zuraski, Wah Chiu, Sarah Cohen, Jan N. Hansen, Manuel D Leonetti, Chad Liu, Emma Lundberg, and Serena Yeung-Levy. Microvqa: A multimodal reasoning benchmark for microscopy-based scientific research. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19552–19564, 2025. 2

[9] I-Hsiang Chen, Wei-Ting Chen, Yu-Wei Liu, Ming-Hsuan Yang, and Sy-Yen Kuo. Improving point-based crowd counting and localization based on auxiliary point guidance. In *European Conference on Computer Vision*, pages 428–444. Springer, 2024. 2

[10] Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024. 2

[11] Etienne David, Simon Madec, Pouria Sadeghi-Tehran, Helge Aasen, Bangyou Zheng, Shouyang Liu, Norbert Kirchgessner, Goro Ishikawa, Koichi Nagasawa, Minhajul A Badhon, et al. Global wheat head detection (gwhd) dataset: A large and diverse dataset of high-resolution rgb-labelled images to develop and benchmark wheat head detection methods. *Plant Phenomics*, 2020. 3

[12] Stanislas Dehaene. *The number sense: How the mind creates mathematics*. Oxford University Press, Oxford, 1997. 1

[13] Yaowu Fan, Jia Wan, Tao Han, Antoni B Chan, and Andy J Ma. Video individual counting for moving drones. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12284–12293, 2025. 1

[14] Lisa Feigenson, Susan Carey, and Marc D. Hauser. Core systems of number. *Trends in Cognitive Sciences*, 8(7):307–314, 2004. 1

[15] Negar Foroutan, Angelika Romanou, Matin Ansaripour, Julian Martin Eisenschlos, Karl Aberer, and Rémi Lebret. Wikimixqa: A multimodal benchmark for question answering over tables and charts. *arXiv preprint arXiv:2506.15594*, 2025. 2

[16] Guangshuai Gao, Junyu Gao, Qingjie Liu, Qi Wang, and Yunhong Wang. A survey of deep learning methods for density estimation and crowd counting. *Vicinagearth*, 2(1):1–37, 2025. 2

[17] Junyu Gao, Qi Wang, and Xuelong Li. Pcc net: Perspective crowd counting via spatial convolutional network. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(10):3486–3498, 2019. 1

[18] Junyu Gao, Liangliang Zhao, and Xuelong Li. Nwpu-moc: A benchmark for fine-grained multicategory object counting in aerial images. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2024. 3

[19] Junyu Gao, Da Zhang, Qiyu Wang, Zhiyuan Zhao, and Xuelong Li. Dynamic proxy domain generalizes the crowd localization by better binary segmentation. *Pattern Recognition*, page 112481, 2025. 1

[20] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017. 2

[21] Taicheng Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh Chawla, Olaf Wiest, Xiangliang Zhang, et al. What can large language models do in chemistry? a comprehensive benchmark on eight tasks. *Advances in Neural Information Processing Systems*, 36:59662–59688, 2023. 1

[22] Tao Han, Lei Bai, Junyu Gao, Qi Wang, and Wanli Ouyang. Dr. vic: Decomposition and reasoning for video individual counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3083–3092, 2022. 1

[23] Tao Han, Lei Bai, Lingbo Liu, and Wanli Ouyang. Steerer: Resolving scale variations for counting and localization via selective inheritance learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21848–21859, 2023. 2

[24] Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv e-prints*, pages arXiv–2507, 2025. 1

[25] Meng-Ru Hsieh, Yen-Liang Lin, and Winston H Hsu. Drone-based object counting by spatially regularized regional proposal network. In *Proceedings of the IEEE international conference on computer vision*, pages 4145–4153, 2017. 2

[26] Yuxuan Hu, Haibin Wu, Ruchao Fan, Xiaofei Wang, Heng Lu, Yao Qian, and Jinyu Li. Phi-omni-st: A multimodal language model for direct speech-to-speech translation. *arXiv preprint arXiv:2506.04392*, 2025. 2

[27] Haroon Idrees, Imran Saleemi, Cody Seibert, and Mubarak Shah. Multi-source multi-scale counting in extremely dense crowd images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2547–2554, 2013. 2

[28] Haroon Idrees, Muhmmad Tayyab, Kishan Athrey, Dong Zhang, Somaya Al-Maadeed, Nasir Rajpoot, and Mubarak Shah. Composition loss for counting, density map estimation and localization in dense crowds. In *Proceedings of the European conference on computer vision (ECCV)*, pages 532–546, 2018. 3

[29] Dimitris Samaras Jingyi Xu, Hieu Le. Lvlm-count: Enhancing the counting ability of large vision-language models. *arXiv preprint arXiv:2412.00686*, 2024. 1

[30] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13299–13308, 2024. 1, 2

[31] Yuhong Li, Xiaofan Zhang, and Deming Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1091–1100, 2018. 2

[32] Dingkang Liang, Wei Xu, and Xiang Bai. An end-to-end transformer model for crowd localization. In *European Conference on Computer Vision*, pages 38–54. Springer, 2022. 2

[33] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024. 1, 2

[34] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, page 2263–2279, 2022. 2

[35] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021. 1, 2

[36] K Fountoulakis MF Qharabagh, M Ghofrani. Zero-shot object counting with language-vision models. *arXiv preprint arXiv:2309.13097*, 2023. 1

[37] Parya Haji Mirzaee, Mohammad Shojafar, Haitham Cruickshank, and Rahim Tafazolli. Chfl: A collaborative hierarchical federated intrusion detection system for vehicular networks. In *2022 IEEE Symposium on Computers and Communications (ISCC)*, pages 1–7. IEEE, 2022. 2

[38] MD Holla MR Holla, D Suma. Optimizing accuracy and efficiency in real-time people counting with cascaded object detection. *International Journal of Information Technology*, 2024. 1

[39] Viresh Ranjan, Udbhav Sharma, Thu Nguyen, and Minh Hoai. Learning to count everything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3394–3403, 2021. 2

[40] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 2

[41] Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. Scienceqa: A novel resource for question answering on scholarly articles. *International Journal on Digital Libraries*, 23(3):289–301, 2022. 1

[42] Deepak Babu Sam, Skand Vishwanath Peri, Mukuntha Narayanan Sundararaman, Amogh Kamath, and R Venkatesh Babu. Locate, size, and count: accurately resolving people in dense crowds via detection. *IEEE transactions on pattern analysis and machine intelligence*, 43(8): 2739–2751, 2020. 2

[43] Thomas Sgouros and Nikolaos Mitianoudis. A novel directional framework for source counting and source separation in instantaneous underdetermined audio mixtures. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2025–2035, 2020. 2

[44] Vishwanath A Sindagi, Rajeev Yasarla, and Vishal M Patel. Jhu-crowd++: Large-scale crowd counting dataset and a benchmark method. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2594–2609, 2020. 3

[45] Qingyu Song, Changan Wang, Zhengkai Jiang, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Yang Wu. Rethinking counting and localization in crowds: A purely point-based framework. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3365–3374, 2021. 2

[46] Guolei Sun, Zhaochong An, Yun Liu, Ce Liu, Christos Sakaridis, Deng-Ping Fan, and Luc Van Gool. Indiscernible object counting in underwater scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13791–13801, 2023. 3

[47] Alexandra Swanson, Margaret Kosmala, Chris Lintott, Robert Simpson, Arfon Smith, and Craig Packer. Snapshot serengeti, high-frequency annotated camera trap images of 40 mammalian species in an african savanna. *Scientific data*, 2(1):1–14, 2015. 3

[48] Alexander Trott, Caiming Xiong, and Richard Socher. Interpretable counting for visual question answering. *arXiv preprint arXiv:1712.08697*, 2017. 2

[49] Nicolas Turpault, Romain Serizel, Ankit Parag Shah, and Justin Salamon. Sound event detection in domestic environments with weakly labeled data and soundscape synthesis. In *Workshop on Detection and Classification of Acoustic Scenes and Events*, 2019. 3, 5

[50] Junfeng Wan, Shuhao Qian, Zihan Tian, and Yanyun Zhao. An effective framework of multi-class product counting and

recognition for automated retail checkout. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3282–3290, 2022. 1

[51] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169, 2024. 1

[52] Qi Wang, Junyu Gao, Wei Lin, and Yuan Yuan. Learning from synthetic data for crowd counting in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8198–8207, 2019. 2

[53] Qi Wang, Junyu Gao, Wei Lin, and Xuelong Li. Nwpu-crowd: A large-scale benchmark for crowd counting and localization. *IEEE transactions on pattern analysis and machine intelligence*, 43(6):2141–2149, 2020. 2

[54] 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*, 2025. 1

[55] Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. *arXiv preprint arXiv:2307.10635*, 2023. 1

[56] Yibo Wang, Wenting Zhao, Yao Wan, Zhongfen Deng, and Philip S Yu. Named entity recognition via machine reading comprehension: A multi-task learning approach. *arXiv preprint arXiv:2309.11027*, 2023. 1

[57] 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*, 2025. 2

[58] Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025. 2

[59] Fan Yu, Shiliang Zhang, Yihui Fu, Lei Xie, Siqi Zheng, Zhihao Du, Weilong Huang, Pengcheng Guo, Zhijie Yan, Bin Ma, Xin Xu, and Hui Bu. M2MeT: The ICASSP 2022 multi-channel multi-party meeting transcription challenge. In *Proc. ICASSP*. IEEE, 2022. 3, 5

[60] Yuan Yuan, Haojie Guo, and Junyu Gao. Distance-aware network for physical-world object distribution estimation and counting. *Pattern Recognition*, 157:110896, 2025. 1

[61] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024. 1, 2

[62] Da Zhang, Feiyu Wang, Bingyu Li, Zhiyuan Zhao, Junyu Gao, and Xuelong Li. Kaid: Knowledge-aware interactive distillation for vision-language models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 3212–3221, 2025. 1

[63] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 589–597, 2016. 2

[64] Yan Zhang, Jonathon Hare, and Adam Prügel-Bennett. Learning to count objects in natural images for visual question answering. *arXiv preprint arXiv:1802.05766*, 2018. 2

[65] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023. 3

[66] Henry Hengyuan Zhao, Pan Zhou, Difei Gao, Zechen Bai, and Mike Zheng Shou. Lova3: Learning to visual question answering, asking and assessment. *Advances in Neural Information Processing Systems*, 37:115146–115175, 2024. 1