UNICBench: UNIfied Counting Benchmark for MLLM

Supplementary Material

6. Benchmark Construction Details

6.1. Label Definition

All modalities share a unified JSON structure with common fields (`type`, `target_file_path`, `attributes`, `questions`) and modality-specific extensions. Each question object contains bilingual descriptions, difficulty level, ground truth count, and an `instances` array. Instance annotations adapt to modality characteristics: images use (x, y) pixel coordinates, text employs $[\text{start}, \text{end})$ character indices with matched text snippets, and audio records sound events with temporal ranges $[t_{\text{start}}, t_{\text{end}}]$ in seconds. This design enables consistent cross-modal evaluation while preserving modality-specific localization granularity.

The complete JSON schema is shown below:

```
{
  "type": "image/text/audio",
  "target_file_path": "filename or false (for
      text)",
  "attributes": {
    "height": "int (image only)",
    "width": "int (image only)",
    "duration": "float (audio only, in seconds)",
    "format": "string (audio only)",
    "ori_file": "string (text only)",
    "sub_type": "string (text optional)"
  },
  "target_text": "string (text only)",
  "questions": [
    {
      "question_id": "int",
      "question": "English description",
      "question_cn": "Chinese description",
      "level": "Pattern/Semantic/Reasoning",
      "count": "int",
      "instances": [
        "Image: [x, y]",
        "Text: {text: ..., coordinates: {start:
            int, end: int}}",
        "Audio: {sound_type: ..., time_range: [
            start, end]}"
      ]
    }
  ]
}
```

6.2. Construction details for different modality

Image Modality Construction. The 5,300 image samples span 49 categories with varying source compositions. From FSC147, we extract 30+ everyday object categories (apples, beads, birds, biscuits, books, bottle caps, chairs, cups, etc.), each contributing 100 samples selected for count range diversity. NWPU-MOC provides remote sensing imagery (airplane, boat, farmland, house, industrial, mansion, pool, stadium, tree, truck) where objects exhibit scale variation and occlusion. Crowd counting integrates 150 samples from JHU-CROWD++, UCF-QNRF, ShanghaiTech Part A, and NWPU-Crowd, with manual supplementation for extreme density scenarios (average count: 355.58, max: 10,294). The medical domain contributes 400 cell images from Blood Cell Count and Detection Dataset (average: 50.06 cells/image). Notably, the screen panel category (50 samples, 55 questions) involves complete team annotation using custom labeling tools, while pens, birds, seagulls, books, biscuits, chairs, donuts tray, marbles, cups, mini blinds, potatoes, alcohol bottles, crows, green peas, and lipstick receive manual enhancement beyond FSC147's base annotations. Car samples merge CARPK and NWPU-MOC (100 each) to ensure scene diversity. Question design follows a stratified protocol: 99.1% Pattern-level for direct visual counting, 0.5% Semantic-level for attribute-specific queries, and 0.4% Reasoning-level for relational constraints. Annotation verification employs cross-checking against original dataset ground truths, with point annotations converted to instance centers via bounding box centroids or density map peaks.

Text Modality Construction. The 872 text samples yielding 5,888 questions represent entirely self-collected data across 12 categories. Code samples (100 Java/Python files, 500 questions) target structural elements like functions, classes, and loops (average count: 12.21). JSON data (200 samples, 1,800 questions) emphasizes nested structures and key-value pairs (average: 58.87, max: 71,176 in LaTeX category with 66 samples and 593 questions due to citation and formula density). HTML samples (168 files, 1,176 questions) focus on tag hierarchies and attributes (average: 3.54), where each page is compressed into a content-preserving fragment by stripping redundant elements (e.g., embedded base64 images, SVG/Canvas objects) while retaining core structural nodes. Musical Notation leverages MusicXML format (143 files, 715 questions) to count notes, measures, and rests (average: 108.58). Ancient texts (20 samples, 96 questions) and Literary works (60 samples, 109 questions) use language-preserving questions matching source text style—classical Chinese for ancient documents and original language for literary texts. CSV data (15 files, 150 questions) involves large-scale tabular counting (average: 291.3). Legal documents (10 samples, 105 questions), Exam papers (15 samples, 120 questions), News articles (60 samples, 300 questions), and Official documents (15 samples, 224 questions) round out the corpus. Question distribution skews toward higher complexity: 26.4% Pattern-

Table 5. Modality-Specific Construction Statistics

Modality	Samples	Questions	Categories	Avg Count	Count Range
Image	5,300	5,508	49	63.53	[0, 10,294]
Text	872	5,888	12	120.40	[0, 71,176]
Audio	2,069	2,905	2	30.32	[0, 635]
Total	**8,241**	**14,301**	**63**	**71.42**	**[0, 71,176]**

level, 43.7% Reasoning-level, and 29.9% Semantic-level. Text length varies dramatically (584 to 8,052,974 characters, median: 7,176), requiring models to handle both short snippets and long documents. Annotators employ automated pre-counting tools (regex, syntax parsers) followed by manual verification, with each sample receiving dual annotation for counts exceeding 50 instances.

Audio Modality Construction. The 2,069 audio samples derive from two specialized datasets. DESED contributes 1,860 environmental sound samples (doors, alarms, dog barks, keyboard typing) with event-level timestamps, exhibiting sparse event density (average: 1.56 events/sample, duration: 3–300 seconds). AliMeeting provides 209 meeting recordings (1,045 questions) with dense speech segments (average: 81.51 counts/sample, duration: 10–2,497.9 seconds, median: 10.0). Temporal annotations achieve frame-level precision (0.01-second granularity), with sound_type labels distinguishing speaker identities ("Speaker1_unknown"), event categories ("Cat meowing"), and semantic units ("Question"). Audio preprocessing standardizes WAV format with original sampling rates preserved, applying minimal noise reduction to retain naturalistic characteristics. Question design balances perceptual tasks (64.0% Pattern-level: counting discrete events) with semantic challenges (36.0% Semantic/Reasoning-level: speaker identification, turn-taking analysis).

6.3. More Data Statistics

We provide more detailed statistics to further reveal the distributional characteristics of our dataset across the three major modalities—**Image, Text, and Audio**. Table 5 summarizes the overall construction statistics for each modality, including sample counts, question counts, category coverage, and count-value ranges. In addition, Figure 7 provides a comprehensive cross-modal comparison, visualizing key differences in sample composition, average count values, difficulty distributions, and count-scale variability across the three modalities. Together, these statistics outline the macro-level properties of UNICBench and provide context for the finer-grained analyses shown in Figures 8–11.

These analyses highlight the broad coverage and intrinsic difficulty of UNICBench. Across image, text, and audio modalities, the dataset spans diverse count ranges, distinct question difficulty patterns, and large variations in question length and structural complexity. Category-level

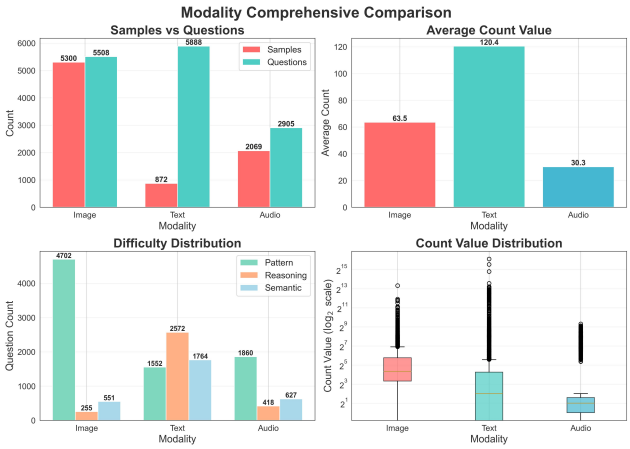


Figure 7. **Modality comprehensive comparison.** The figure illustrates cross-modal distributions among image, text, and audio, comparing sample and question counts, average values, difficulty types, and count distributions.

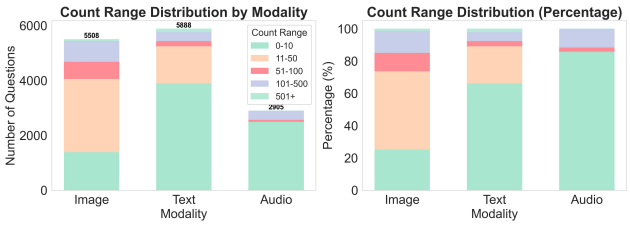


Figure 8. **Count-range distribution across the three modalities.** Image, text, and audio samples exhibit very different count scales, where text spans the widest range while image and audio concentrate in lower ranges—reflecting distinct counting difficulty patterns across modalities.

statistics further reveal substantial intra-modality heterogeneity—from high-density structural counts in text to perceptual pattern counts in images and temporally grounded counts in audio. Together, these results show that our benchmark captures the full spectrum of real-world counting scenarios, providing a unified and challenging evaluation suite for large multimodal models.

6.4. Cross-Difficulty Comparison

Table 6 summarizes the difficulty progression across modalities:

Table 6. Sample Difficulty Characteristics

Level	Cognitive Demand	Example Tasks
Pattern (L1)	Direct perception	Count visible objects, discrete events
Semantic (L2)	Attribute filtering + deduplication	Unique libraries, distinct speakers
Reasoning (L3)	Multi-step inference	Rule-based counting, logical constraints

These difficulty levels characterize the cognitive gradient of counting in UNICBench: from basic perceptual counting (L1), to attribute-constrained and deduplicated conditional counting (L2), and finally to multi-step reasoning counts

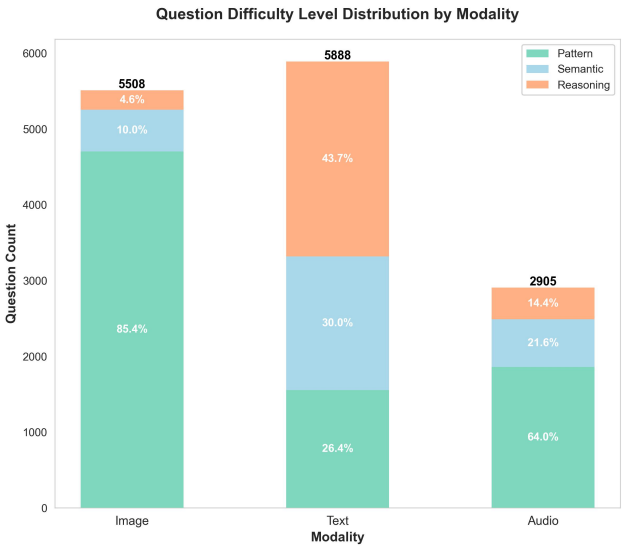


Figure 9. **Distribution of question difficulty levels across modalities.** Image questions are predominantly Pattern-level, whereas Text contains substantially more Reasoning and Semantic questions. Audio shows a more balanced mix, highlighting modality-dependent complexity differences.

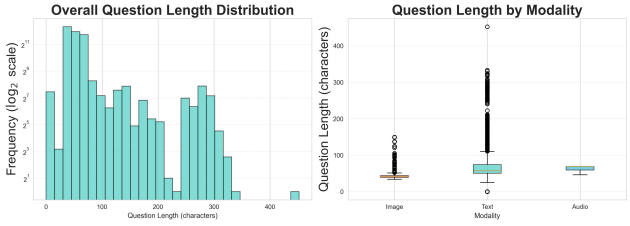


Figure 10. **Question length distribution across modalities.** The left histogram shows the overall length distribution of all counting questions, while the right boxplot compares question lengths across image, text, and audio modalities. Text questions are significantly longer and more variable, indicating higher linguistic complexity and reflecting the greater reasoning demand in text-based counting tasks.

(L3). This hierarchy provides a clear, controlled, and interpretable difficulty structure for cross-modal counting evaluation.

6.5. Data & Result Samples

This section presents representative examples from each modality to illustrate the diversity and complexity of counting tasks in **UNICBench**. Each example showcases the input format, counting questions and expected model outputs across different difficulty levels.

6.5.1. Image Modality

We select three representative image samples corresponding to the three difficulty levels (L1–L3) and present the counting results of several models. These examples illustrate the progression of visual counting difficulty—from simple perceptual enumeration to attribute-constrained and reasoning-based scenarios—and highlight both the characteristic patterns of our dataset and the capability differences across models, as shown in Figures 12–14.

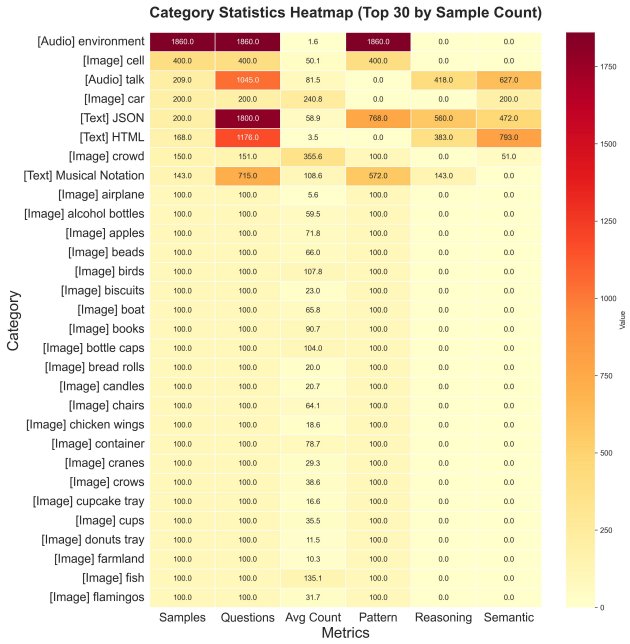


Figure 11. **Category-level statistics heatmap (top 30 by sample count).** This heatmap summarizes the top 30 categories by sample count, showing clear modality-dependent differences: audio–environment contributes large volumes with low counts, text categories exhibit high-density and complex count structures, while image categories remain more uniform with moderate difficulty.

6.5.2. Text Modality

For the text modality, we also present three representative examples spanning levels L1–L3, together with predictions from various models. These cases show how counting in text shifts from surface-level token matching to attribute-based filtering and multi-step reasoning over long-range dependencies, revealing modality-specific challenges that complement the visual examples in Figures 15–17.

6.5.3. Audio Modality

In the audio modality, we also present three representative examples covering levels L1–L3. These samples illustrate how counting in audio is shaped by acoustic cues such as rhythmic repetition, temporal sparsity, overlapping events, and variability in signal quality—factors that impose challenges distinct from those in vision or text. Representative clips for each difficulty level, along with model predictions, are shown in Figures 18–20.

Level: Pattern (L1)
Size: 520×384 pixels
Category: lipstick
Ground Truth Count: 50

System Prompt: 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.
Question: How many lipstick are there in the image?

: 72 GPT-4o | : 35 Gemini 2.5 Pro | : 40 Claude 4 | : 30 Qwen3-VL-30B-A3B

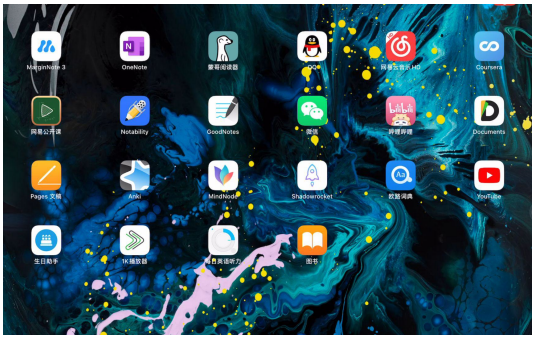
Figure 12. Example of a pattern-level sample from the image modality and the answer from different models.

Level: Semantic (L2)
Size: 750×563 pixels
Category: crowd
Ground Truth Count: 22

System Prompt: 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.
Question: How many real people are directly visible in the image?

: 34 GPT-4o | : 37 Gemini 2.5 Pro | : 32 Claude 4 | : 30 Qwen3-VL-30B-A3B

Figure 13. Example of a Semantic-level sample from the image modality,and the answer from different models.



Level: Reasoning (L3)
Size: 1716×1072 pixels
Category: screen panel
Ground Truth Count: 22

System Prompt: 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.
Question: How many applications are there in the desktop screenshot?

: 24 GPT-4o | : 22 Gemini 2.5 Pro | : 22 Claude 4 | : 20 Qwen3-VL-30B-A3B

Figure 14. Example of a reasoning-level sample from the image modality, and the answer from different models.

MusicXML Score Snippet

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE score-partwise PUBLIC
"-//Recordare//DTD MusicXML 3.1
Partwise//EN"
"http://www.musicxml.org/partwise.dtd">
<score-partwise version="3.1">
<work>
<work-title>Meine Seele erhebet den
Herrn</work-title>
</work>
<movement-title>Meine Seele erhebet den
Herrn</movement-title>
<identification>
<creator type="composer">Bach, Johann
Sebastian</creator>...
```

Level: Pattern (L1)
Length: 100,913 characters (4,684 words)
Category: XML document
Ground Truth Answer: 272

System Prompt: You are a professional text analysis and counting expert. You MUST answer ALL questions, do NOT skip any. You will receive 9 questions and MUST provide exactly 9 numeric answers. NEVER provide fewer or more answers than required.
Question(2): What is the total number of note nodes in the entire piece?

: 274 GPT-5 | : 275 Gemini 2.5 Pro | : 20 DeepSeek-R1-0528 | : 267 GLM-4.6

Figure 15. Example of a pattern-level sample from the text modality, and responses from different models.

《过秦论·上篇》—— 贾谊（西汉）

秦孝公据崤函之固，拥雍州之地，君臣固守以窥周室，有席卷天下、包举宇内、囊括四海之意，并吞八荒之心。商君佐之，内立法度，务耕织，修守战之具，外连衡而斗诸侯，于是秦人拱手而取西河之外。

惠文、武、昭襄因遗策，南取汉中，西举巴、蜀，东割膏腴之地，北收要害之郡。诸侯恐惧，会盟而谋弱秦，不爱珍器重宝、肥饶之地，以致天下之士，合从缔交，相与为一。

齐有孟尝，赵有平原，楚有春申，魏有信陵，皆明智忠信，尊贤重士，约从离衡，并韩、魏、燕、楚、齐、赵、宋、卫、中山之众。吴起、孙膑、乐毅、廉颇、赵奢之伦制其兵，九国之师攻秦，而秦开关延敌，诸侯逡巡不敢进。

秦无亡矢遗镞之费，而诸侯已困。强国请服，弱国...

Level: Semantic (L2)
Length: 2,867 characters (2,342 words)
Category: Ancient
Ground Truth Answer: 56

System Prompt: 你是一个专业的文本分析计数专家。你必须回答所有问题，不能跳过任何一个。你将收到5个问题，必须给出恰好5个数字答案。绝不能少于或多于要求的答案数量。
Question(1): 全文中明写的数字（汉字或阿拉伯数字）共有多少处?

: 56 GPT-5 | : 47 Gemini 2.5 Pro | : 53 DeepSeek-R1-0528 | : 55 GLM-4.6

Figure 16. Example of a semantic-level sample from the text modality, and responses from different models.

***Heart of Darkness* (Excerpt)**

"Next day I left that station at last, with a caravan of sixty men, for a two-hundred-mile tramp.

No use telling you much about that. Paths, paths, everywhere; a stamped-in network of paths spreading over the empty land, through the long grass, through burnt grass, through thickets, down and up chilly ravines, up and down stony hills ablaze with heat; and a solitude, a solitude, nobody, not a hut. The population had cleared out a long time ago. Still I passed through several abandoned villages. Day after day, with the stamp and shuffle of sixty pair of bare feet behind me, each pair under a 60-lb. load.

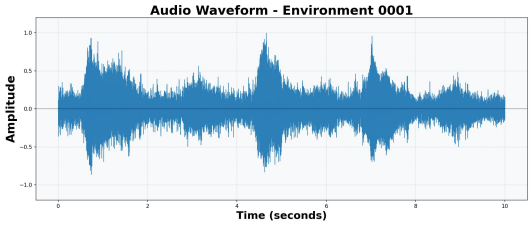
Now and then a carrier dead in harness, at rest in the long grass near the path, with an empty water-gourd and his long staff lying by his side. Once a white man in an unbuttoned uniform, camping on the path with an armed escort of lank...

Level: Reasoning (L3)
Length: 46,391 characters (8,399 words)
Category: literary narrative
Ground Truth Answer: 5

System Prompt: You are a professional text analysis and counting expert. You MUST answer ALL questions, do NOT skip any. You will receive 1 questions and MUST provide exactly 1 numeric answers. NEVER provide fewer or more answers than required.
Question: Before Marlow reaches the big river, how many separate mishaps or disasters does he recount?

: 5 GPT-5 | : 6 Gemini 2.5 Pro | : 5 DeepSeek-R1-0528 | : 5 GLM-4.6

Figure 17. Example of a reasoning-level sample from the text modality, and responses from different models.

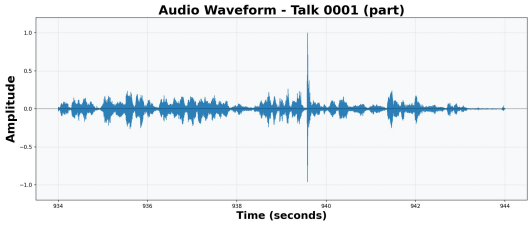


Level: Pattern (L1)
Size: 10 S
Category: environment
Ground Truth Count: 1

System Prompt: 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.
Question: In this audio, how many different types of sounds are there in total?

: 2 GPT-4o-Audio-Preview | : 3 Gemini 2.5 Pro | : 2 Phi-4-Multimodal | : 2 Qwen3-Omni-30B-A3B

Figure 18. Example of a pattern-level sample from the audio modality and the answer from different models.

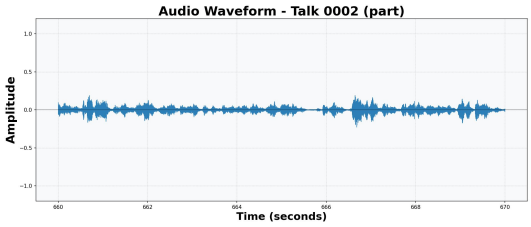


Level: Semantic (L2)
Size: 2040.69 S
Category: talk
Ground Truth Count: 3

System Prompt: 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.
Question: In this audio, how many female speakers are there in total?

: failed GPT-4o-Audio-Preview | : failed Gemini 2.5 Pro | : failed Phi-4-Multimodal | : 6 Qwen3-Omni-30B-A3B

Figure 19. Example of a semantic-level sample from the audio modality and the answer from different models.



Level: Reasoning (L3)
Size: 2129.05 S
Category: talk
Ground Truth Count: 149

System Prompt: 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.
Question: In this audio, how many questions are asked in total?

: failed GPT-4o-Audio-Preview | : failed Gemini 2.5 Pro | : failed Phi-4-Multimodal | : 1 Qwen3-Omni-30B-A3B

Figure 20. Example of a reasoning-level sample from the audio modality and the answer from different models.

7. Experiment Results and Analysis

7.1. More Settings

To enable efficient large-scale evaluation, we feed all questions associated with the same document to the model in a single batch, and require the model to answer them strictly in the order they are asked. This batching strategy significantly accelerates inference while maintaining fair comparison among different models.

Because questions are processed together, we slightly modify our system prompts and user prompts to explicitly constrain the model's behavior and ensure sequential answering. Both prompts are provided in Chinese and English versions, and we select the appropriate version according to the language of the underlying document. The English versions of the system prompt and user prompt are as follows:

English System Prompt

You are a professional text analysis and counting expert. You MUST answer ALL questions, do NOT skip any.
You will receive {len(questions)} questions and MUST provide exactly {len(questions)} numeric answers.
NEVER provide fewer or more answers than required.

English User Prompt

Please carefully read the following text and answer all questions in order.
Text Content: {`text_display`}
Please answer the following {len(questions)} questions in order, providing only a number for each:
`Q1`
`Q2`
`...`
`Qn`
STRICT REQUIREMENTS:
- NEVER output any analysis process or explanations.
- You MUST answer ALL questions.
- Output exactly {len(questions)} numeric answers.
- Separate answers with commas.

Example format(5 answers): `5,23,0,17,8`
Your answer:

Thinking Models. These modes are toggled by configuration parameters (e.g., enable_thinking:True, thinking:True).

7.2. More Results of Three Modality

To provide a more comprehensive view of model behavior across different input types, we report additional experimental results for the image, text, and audio modalities. These analyses complement the main paper by highlighting modality-specific characteristics in counting performance—ranging from visually grounded object enumeration, to structure-dependent textual counting, and temporally distributed auditory event counting. For each modality, we present accuracy trends and difficulty-level comparisons, offering a deeper understanding of how current MLLMs handle numerosity under different perceptual and reasoning demands.

Image Results. The image modality reflects the most classical setting for visual counting, where models must infer numerosity from complex scenes containing varying object scales, dense crowds, occlusions, and visual clutter—all without any requirement for explicit localization or detection outputs. Instead, models are evaluated solely on their ability to produce the correct count. To examine how current MLLMs handle these challenges, we report two complementary results: overall counting accuracy across models and performance breakdown by difficulty level.

Figure 21 presents the accuracy comparison under Exact Match and relaxed error tolerances (10% and 20%). This figure highlights substantial performance variation among models, with Exact Match accuracy remaining low overall—indicating that precise object enumeration in complex visual scenes is still challenging. Allowing small error tolerance significantly improves accuracy, showing that models often produce approximate but not exact counts.

Figure 22 further analyzes model performance across the three difficulty levels (Pattern, Semantic, Reasoning). The results show a clear difficulty gradient: models perform best on perceptual Pattern-level cases, moderately well on attribute-dependent Semantic samples, and experience the largest degradation on Reasoning-level scenes requiring multi-step constraints or cross-region aggregation. The consistent decline across levels underscores that visual reasoning, rather than perception alone, is the primary bottleneck for visual counting.

Text Results. The text modality poses a different type of counting challenge compared to images: models must infer numerosity from text sequences and document structures rather than from visual patterns. Many of our text tasks require understanding lists, tables, code blocks, or long-form prose, and often involve operations such as filtering, grouping, or aggregating entities described in natural language. Since we only evaluate the final numeric answer, models are expected to perform this reasoning implicitly over diverse textual formats. To characterize their behavior, we report overall counting accuracy across models and a breakdown of performance by difficulty level.

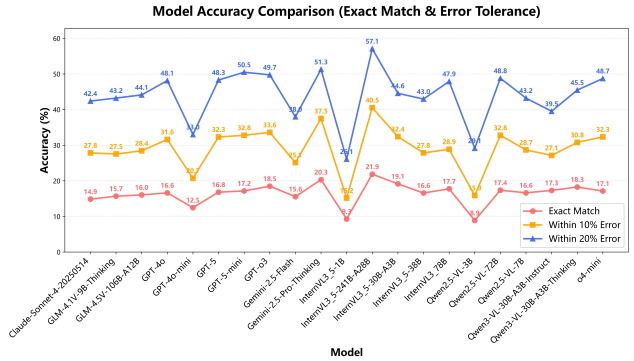


Figure 21. **Image modality accuracy comparison.** Exact-match accuracy remains low across models, while allowing small error tolerance (10% and 20%) yields significantly higher performance, revealing strong approximate-counting ability but limited precise counting.

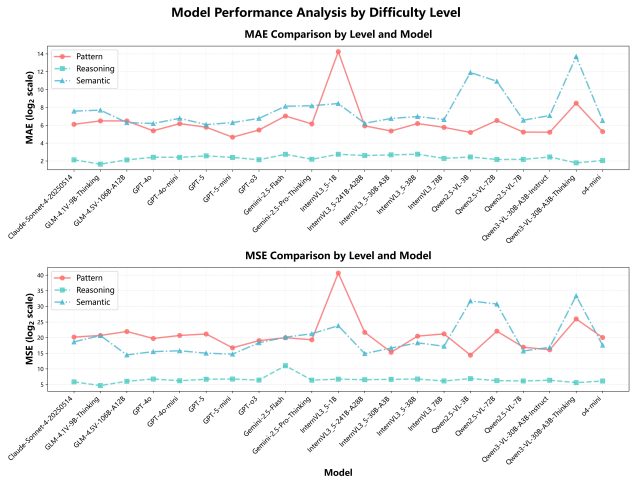


Figure 22. **Performance across difficulty levels in the image modality.** Models perform best on Pattern-level (L1) samples, degrade on Semantic-level (L2), and show the largest errors on Reasoning-level (L3) samples, demonstrating the expected progression of counting difficulty.

Figure 23 compares text-counting accuracy across models under three criteria: Exact Match, within 10% relative error, and within 20% relative error. Overall, larger models achieve noticeably higher accuracies, and allowing a small error tolerance substantially boosts performance, indicating that most models can produce roughly correct counts but still struggle with exact numerosity in complex textual contexts.

Figure 24 further breaks down performance across the three difficulty levels (L1–L3). We observe a clear degradation from Pattern-level tasks to Semantic- and Reasoning-level tasks, showing that counting becomes harder when models must incorporate attribute constraints, resolve references, or perform multi-step reasoning over longer spans of text.

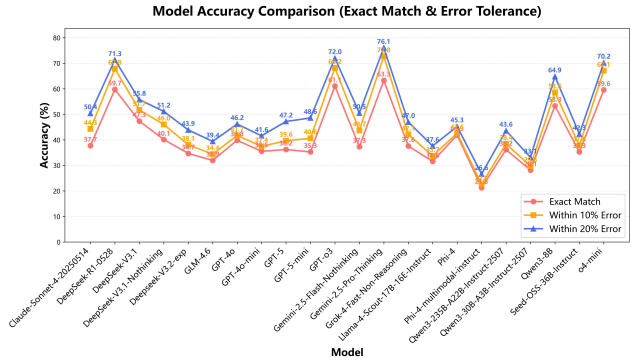


Figure 23. **Text modality accuracy comparison.** Models achieve higher accuracy on text counting than on other modalities under relaxed error thresholds, but Exact Match remains challenging, especially for complex or long documents.

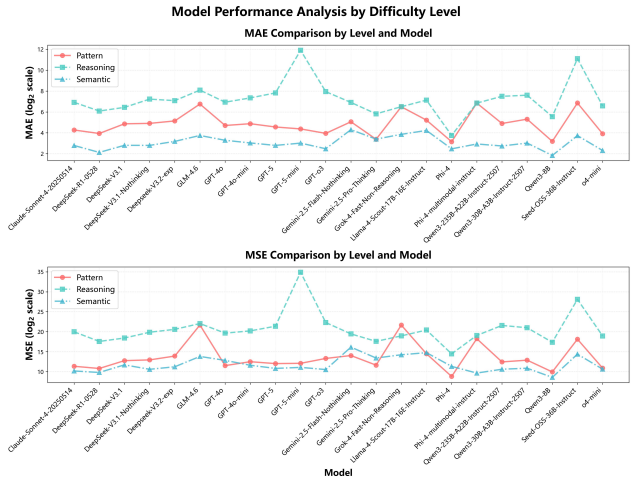


Figure 24. **Performance across difficulty levels in the text modality.** Performance degrades from Pattern-level (L1) to Semantic-level (L2) and Reasoning-level (L3), reflecting the increased complexity of attribute filtering, reference resolution, and multi-step reasoning required by higher levels.

Audio Results. The audio modality introduces a distinct form of counting challenge, where models must infer numerosity purely from temporal acoustic cues rather than spatial or textual structure. Counting in audio depends on detecting discrete events—such as speaker turns, syllabic patterns, or repetitive sounds—distributed across time, often with variations in volume, rhythm, and background noise. To understand how current MLLMs handle these temporal counting tasks, we present two complementary analyses: Overall accuracy under different error tolerances, and performance across the three difficulty levels defined in UNICBench.

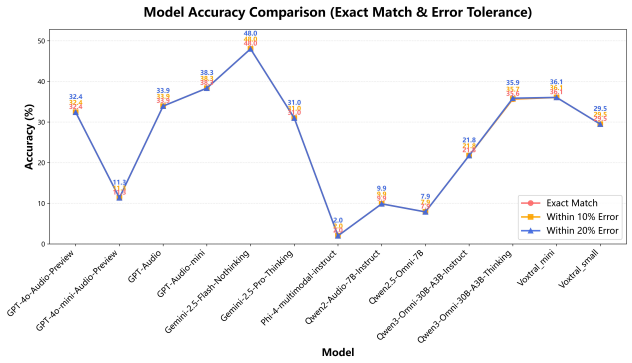


Figure 25. **Audio modality accuracy comparison.** Overall accuracy of audio-capable models under Exact Match, 10% error, and 20% error thresholds. Audio counting shows lower precision due to temporal ambiguity and variable acoustic patterns.

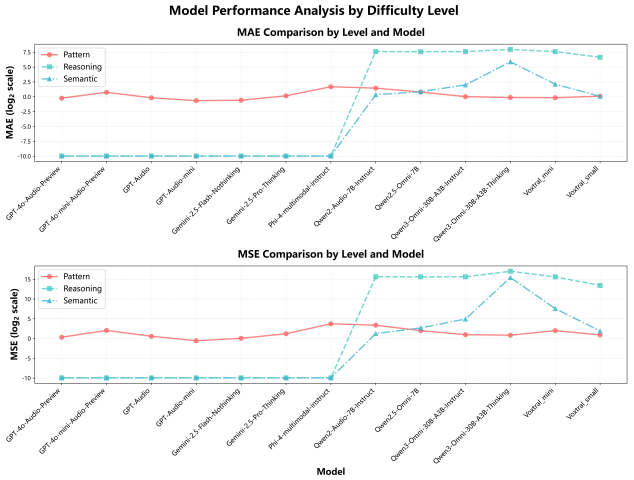


Figure 26. **Audio performance across difficulty levels.** MAE and MSE ($\log_2$ scale) across L1–L3 tasks show increasing error with complexity, highlighting the challenge of temporal reasoning in audio-based counting.

Figure 25 reports accuracy across all evaluated audio-capable models. A notable difference from the image and text modalities is that the three metrics—Exact Match, within 10% error, and within 20% error—show almost no separation. This is because, for all valid responses, the ground-truth counts in the audio modality are relatively small; once a model predicts an incorrect number, even a 20% tolerance is insufficient to bring the answer into the acceptable range. As a result, relaxed thresholds offer little improvement, revealing that model errors are typically categorical (e.g., missing or hallucinating events) rather than minor numerical deviations.

Figure 26 breaks down performance across L1–L3 difficulty levels. As the difficulty increases from simple rhythmic patterns (L1) to attribute-conditioned counting (L2) and multi-step temporal reasoning (L3), we observe a clear degradation in performance. L1 is handled relatively well by most models, but L2 and especially L3 introduce significant error increases. This trend indicates that current MLLMs struggle not only with fine-grained event detection but also with higher-level reasoning over complex auditory sequences.

7.3. Error Analysis

Before analyzing modality-specific counting errors, we first examine the failure modes that emerge during large-scale evaluation. Although models are instructed to respond with a single numeric answer, a non-negligible portion of responses deviate from the required format or fail to produce valid counts. These errors are grouped into three categories:

- **Out-of-Context**: The model generates content unrelated to the question, often due to excessive context length, insufficient attention to the query, or internal prioritization of irrelevant cues. In the audio modality, this category also includes cases where inputs exceed the 20 MB file-size limit of the Azure platform, causing the model to fail before processing the actual content.
- **Out-of-Thinking**: Long or incomplete "thinking" traces interfere with template extraction. This often occurs when models expose raw internal reasoning (e.g., extremely long chains), or when API settings do not suppress extended reasoning outputs.
- **Incorrect Format**: The model returns text that does not contain a valid number, such as "I cannot count…" or safety-triggered responses. These outputs cannot be parsed as numeric predictions and are therefore excluded from accuracy calculations.

Table 7. **Distribution of error types across modalities.**

Modality	None	Out of Context	Out of Thinking	Incorrect Format
Image	114,534	0	1,052	82
Text	125,919	7,019	0	2,386
Audio	28,850	6,271	769	830
Total	269,303	13,290	1,821	3,298

Table 7 summarizes the frequency of each error type across the three modalities. Text exhibits the highest rate of formatting-related failures, likely due to longer prompts and more linguistically complex question structures. Audio models show only minor formatting errors but often underprovide responses due to limited audio-counting capability. Image models produce the fewest malformed outputs, reflecting their relatively stable prompt structure under visual counting settings.

These failure patterns highlight that counting errors arise not only from incorrect estimation but also from systemic issues in output formatting and reasoning stability. After removing invalid responses, we conduct a focused analysis on true counting errors, organized by modality. The follow-

ing sections examine how image, text, and audio inputs trigger different forms of numerical deviations and error magnitudes.

7.3.1. Image Modality

Image-based counting, as the most classical form of visual numerosity estimation, reveals some of the most distinct error behaviors in current MLLMs. Although models consistently attempt to provide numerical answers, their predictions can diverge substantially from ground truth due to category-specific difficulty, visual clutter, occlusion, and intrinsic model biases. To characterize these deviations, we examine error patterns both at the model level—capturing the overall magnitude of numerical error—and at the category level, where systematic weaknesses emerge in scenes with dense objects, repetitive structures, or extreme scale variation.

Figure 27 reports model-level MAE and MSE values ($\log_2$ scale), illustrating substantial variation across different MLLMs. While some models maintain moderate numerical deviations, several others exhibit pronounced error spikes, indicating persistent tendencies toward over-counting or under-counting. These differences underscore that stable counting requires not only strong visual recognition ability but also reliable numerical reasoning.

To further diagnose where these errors originate, Figure 28 presents a category–model MAE heatmap. Categories such as *crowd*, *tree*, and *bottle caps* show consistently elevated errors across nearly all models, reflecting their inherent difficulty due to dense layouts, occlusion, or fine-grained repetitive elements. In contrast, categories containing distinct, well-separated instances yield significantly lower errors. Together, these results highlight that image-counting errors arise from a combination of visual scene complexity and model-specific biases in estimating numerosity.

7.3.2. Text Modality

Counting in text requires models to operate beyond surface-level pattern matching and instead perform structured reasoning—identifying relevant spans, filtering attributes, merging duplicated entities, and sometimes executing symbolic-style operations such as list consolidation or template interpretation. These additional cognitive steps introduce unique error sources not present in image or audio modalities. To understand how these challenges affect counting accuracy, we analyze numerical deviations across models and examine how these errors vary across text categories with different structural and semantic demands.

Across models, the MAE/MSE comparison in Figure 29 reveals large variability in numerical deviation, suggesting that text-based counting remains far from solved. Errors escalate particularly for tasks involving long structured documents or heavily nested formats (e.g., LATEX, JSON),

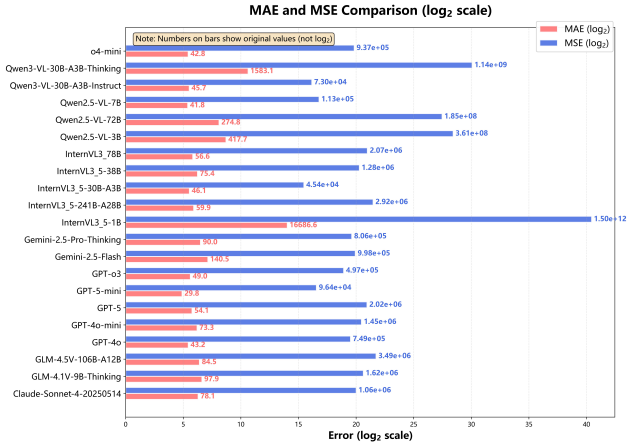


Figure 27. **MAE and MSE comparison for the image modality** ($\log_2$ scale). Numerical labels indicate original (non-log) error values.

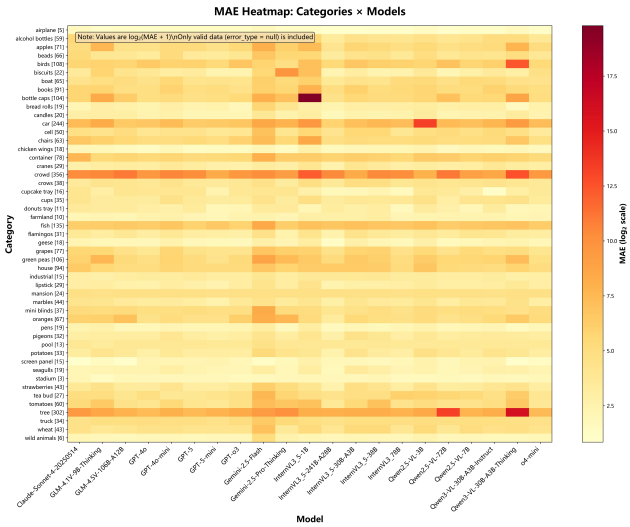


Figure 28. **Category-level MAE heatmap for image-based counting**. Rows denote categories and columns correspond to models. Darker colors represent larger numerical deviations.

where models must correctly parse delimiters and maintain consistency across multi-step reasoning chains.

The category-level heatmap in Figure 30 further highlights this structural sensitivity: categories with rigid syntax (such as LATEX or code) exhibit significantly higher MAE, indicating that even small parsing failures can cascade into large counting mistakes. In contrast, lightweight formats (e.g., short news snippets or CSV-style items) yield relatively lower errors. Overall, these results demonstrate that textual counting difficulty is dominated by reasoning depth and structural complexity rather than document length alone.

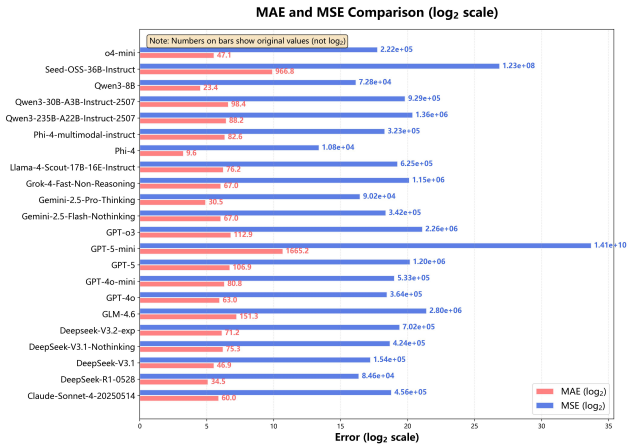


Figure 29. **MAE and MSE comparison for the text modality** ($\log_2$ scale). Numerical labels denote original (non-log) error values.

Figure 30. **Category-level MAE heatmap for text-based counting**. Categories differ widely in structural complexity, causing substantial variation in numerical deviation.

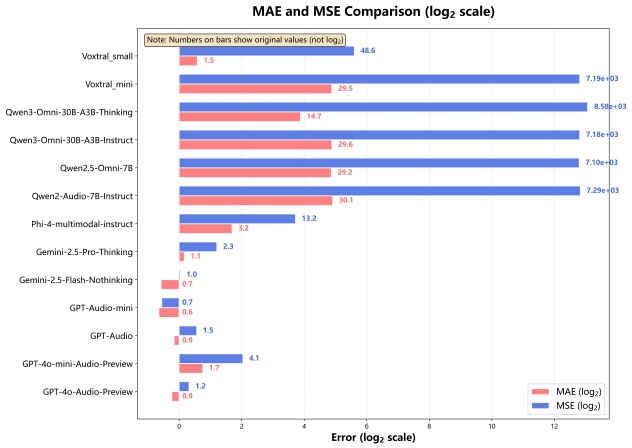


Figure 31. **MAE and MSE comparison for the audio modality** ($\log_2$ scale). Despite small ground-truth counts, numerical deviations remain large across many models.

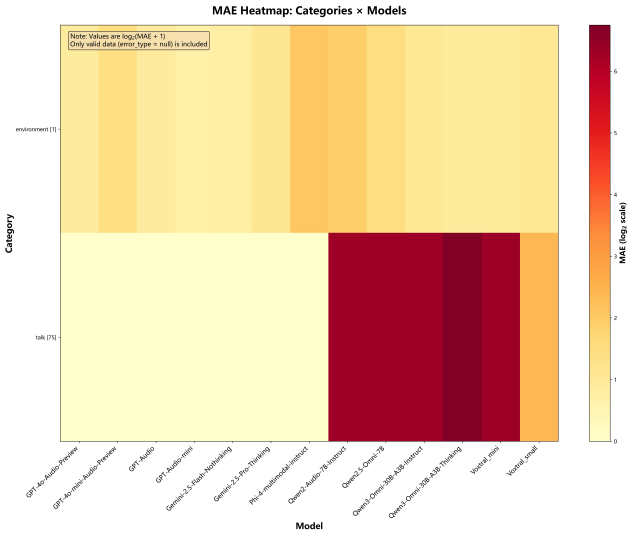


Figure 32. **Category-level MAE heatmap for audio counting**. Environmental sounds are relatively easier, while conversational speech produces disproportionately large errors.

7.3.3. Audio Modality

Counting in the audio modality presents fundamentally different failure modes from visual or textual settings, because temporal signals introduce ambiguity that models cannot easily resolve. Across all evaluated systems, numerical errors remain substantial even when the ground-truth counts are typically small (mostly single-digit). This indicates that mistakes arise not from scale but from the intrinsic difficulty of segmenting acoustic events.

Figure 31 shows that MAE and MSE vary dramatically across models, with several systems exhibiting large deviations even on simple event-counting clips. These errors stem from temporal overlap between events, variable speaking rates, and model sensitivity to background noise. At the category level (Figure 32), we observe a clear performance gap: *environmental* sounds remain relatively manageable for most models, while the *talk* category induces disproportionately large errors, suggesting that conversational audio—with its irregular pauses and overlapping utterances—is significantly harder for current MLLMs to decompose into discrete countable units. Overall, these findings highlight that temporal ambiguity, rather than numerosity magnitude, is the dominant driver of counting errors in audio.