More Images, More Problems? A Controlled Analysis of VLM Failure Modes

Anurag Das[1*], Adrian Bulat[2,3], Alberto Baldrati[2], Ioannis Maniadis Metaxas[2]
Bernt Schiele[1], Georgios Tzimiropoulos[2,4], Brais Martinez[2]
[1]MPI for Informatics, Saarland Informatics Campus [2]Samsung AI, Cambridge
[3]Technical University of Iași, Romania [4]Queen Mary University of London, UK
andas@mpi-inf.mpg.de

Abstract

Large Vision Language Models (LVLMs) have demonstrated remarkable capabilities, yet their ability to understand and reason over multiple images remains largely unexplored. While existing benchmarks have initiated the evaluation of multi-image models, a comprehensive analysis of their core weaknesses and their causes is still lacking. In this work, we introduce MIMIC (Multi-Image Model Insights and Challenges), a new benchmark designed to rigorously evaluate the multi-image capabilities of LVLMs. Using MIMIC, we conduct a series of diagnostic experiments that reveal pervasive issues: LVLMs often fail to aggregate information across images and struggle to track or attend to multiple concepts simultaneously. To address these failures, we propose two novel complementary remedies. On the data side, we present a procedural data-generation strategy that composes single-image annotations into rich, targeted multi-image training examples. On the optimization side, we analyze layer-wise attention patterns and derive an attention-masking scheme tailored for multi-image inputs. Experiments substantially improve cross-image aggregation, while also enhancing performance on existing multi-image benchmarks and outperforming the prior state of the art across tasks. Data and code will be made available at `https://github.com/anurag-198/MIMIC`.

1 Introduction

Current Large Vision Language Models (Li et al., 2025; Liu et al., 2024a; Wang et al., 2024b,a; Yao et al., 2024) showcase impressive vision-language understanding capabilities (Goyal et al., 2017; Mathew et al., 2021; Kembhavi et al., 2016). Most of these models are built upon pre-trained vision encoders (Radford et al., 2021; Zhai et al., 2023) and large language models (LLMs) (Touvron et al., 2023; Abdin et al., 2024; Jiang et al., 2024a). While early efforts primarily focused on single images (Liu et al., 2024a), recent works have extended them to support multiple images (Li et al., 2025; Wang et al., 2024a; Chen et al., 2023) and videos (Zhang et al., 2023) by incorporating temporal modeling and adjusting the positional embeddings (Zhang et al., 2023; Li et al., 2025).

Despite their success, LVLMs continue to face significant challenges (Stogiannidis et al., 2025; Liu et al., 2023; Ouali et al., 2024; Guan et al., 2024; Kaul et al., 2024; Qian et al., 2024). Progress towards identifying and addressing these challenges can follow two primary avenues: the development of comprehensive evaluation benchmarks (Goyal et al., 2017; Mathew et al., 2021; Kembhavi et al., 2016; Masry et al., 2022; Li et al., 2024; Fu et al., 2024a) and the study of the models' inner workings (Deng et al., 2025; Qian et al., 2024; Kaul et al., 2024). To date, research in both areas has predominantly focused on the single-image setting. While early efforts have introduced benchmarks for multi-image scenarios (Wang et al., 2025a; Jiang et al., 2024b; Fu et al., 2024b), a comprehensive, in-depth analysis to ascertain the true efficacy of these models and identify the root causes of their limitations is notably absent.

In this work, we address this gap by conducting a systematic study of LVLMs in multi-image contexts. We first analyze and characterize common failure modes using a newly proposed benchmark, and then seek to mitigate these limitations using two novel complementary fine-tuning strategies. Our in-depth analysis is conducted on the newly introduced **MIMIC** (**M**ulti-**I**mage **M**odel **I**nsights and **C**hallenges) benchmark. Built from MS-COCO (Lin et al., 2014), using its bounding boxes and class labels, MIMIC procedurally generates multi-image sequences by leveraging per-image annotations that give fine-grained control over information spread, distractor presence, object-instance distributions, sequence length, and query complex-

[*]Work done during internship at Samsung AI, Cambridge.

Findings of the Association for Computational Linguistics: ACL 2026, pages 7423–7442
July 2-7, 2026 ©2026 Association for Computational Linguistics

ity, while providing unambiguous ground-truth answers for robust, decorrelated analysis of the model's strengths and weaknesses. Using both quantitative and qualitative assessments, our study reveals that current state-of-the-art LVLMs struggle to effectively aggregate information across multiple images, are unable to track/attend to multiple concepts simultaneously, while being susceptible to distractors. We attribute these shortcomings to a combination of factors, including limitations in multi-image sequence modeling, training data biases, poor inter-image communication induced by the causal attention and the inherent complexity of multi-image reasoning tasks.

Finally, to address the identified problems, we propose two new finetuning strategies: (1) a data-centric approach that generates targeted multi-image training examples to provide rich, multi-image supervision derived from OpenImages (Kuznetsova et al., 2020); and (2) an optimization-centric approach that leverages layer-wise attention analysis to derive an attention-masking scheme tailored for multi-image inputs. Our proposed finetuning strategies lead to substantial performance gains across all scenarios.

In summary, our main contributions are:

- We introduce MIMIC, a comprehensive evaluation framework for multi-image LVLMs that probes various aspects of model performance through a controlled and diverse set of tasks.
- We conduct an extensive evaluation of several state-of-the-art LVLMs using MIMIC, uncovering critical insights into their capabilities and limitations in multi-image settings.
- We propose a novel data-centric finetuning approach using synthetically generated multi-image data, alongside an optimization-centric attention-masking strategy, both of which significantly enhance model performance in multi-image contexts.
- We set new state-of-the-art results on existing multi-image benchmarks, demonstrating the effectiveness of our proposed methods.

2 Related work

Multi-Image Large Vision Language Models: Early LVLMs such as Flamingo (Alayrac et al., 2022) and PaLM-E (Driess et al., 2023) pioneered the integration of pre-trained vision encoders with powerful LLMs for VQA and captioning. Subsequent models (Dai et al., 2023; Li et al., 2025) introduced expanded instruction tuning and multi-modal pre-training techniques. More recent advancements include MiniGPT-5 (Zheng et al., 2024), Qwen2-VL (Wang et al., 2024a), CogVLM2 (Hong et al., 2024) and InternVL3 (Zhu et al., 2025) have further advanced the field by scaling training data and model capacity and adopting more sophisticated architectural designs. While early LVLMs primarily operated on low-resolution, single-image inputs (Liu et al., 2024a; Lin et al., 2024), later research significantly expanded their scope. High-resolution images (Li et al., 2025; Wang et al., 2024a; Zheng et al., 2024) are commonly processed by splitting them into fixed-resolution patches and treating them as image sequences, while videos are represented by extracting frames to form multi-image inputs. In addition, models have begun to explicitly support multi-image context (Wang et al., 2024a; Li et al., 2025; Zhu et al., 2025), enabling reasoning across multiple visual inputs. Multi-image capability is introduced by fine-tuning single-image LVLMs on multi-image instruction-tuning data, while largely preserving the original model architecture and attention mechanisms.

Evaluation of LVLMs: Early evaluation efforts focused on narrower domains with benchmarks such as MS-COCO (Lin et al., 2014), VQA (Antol et al., 2015), DocVQA (Mathew et al., 2021), GQA (Hudson and Manning, 2019) and AI2D (Kembhavi et al., 2016), primarily assessing single-image understanding and using templetized questions with limited diversity. Later work introduced more comprehensive benchmarks to evaluate a wider range of skills, e.g. SEED-Bench (Li et al., 2024), MMBench (Liu et al., 2024c) and MME (Fu et al., 2024a), which feature diverse question types and require complex reasoning abilities. Similarly, video benchmarks such as MMVU (Zhao et al., 2025) and VideoMME (Fu et al., 2025) require models to understand temporal dynamics and to reason across multiple frames.

Closer to our work, several benchmarks have been proposed specifically for multi-image LVLMs. MuirBench (Wang et al., 2025a) introduced 12 tasks evaluating multi-image understanding, including image comparison and multi-image reasoning. Blink (Fu et al., 2024b) included 14 tasks deemed easy for humans, highlighting LVLMs' limitations in truly understanding multi-image visual content. Visual Haystack (Wu et al., 2025) focuses on retrieval-based tasks, assessing how well

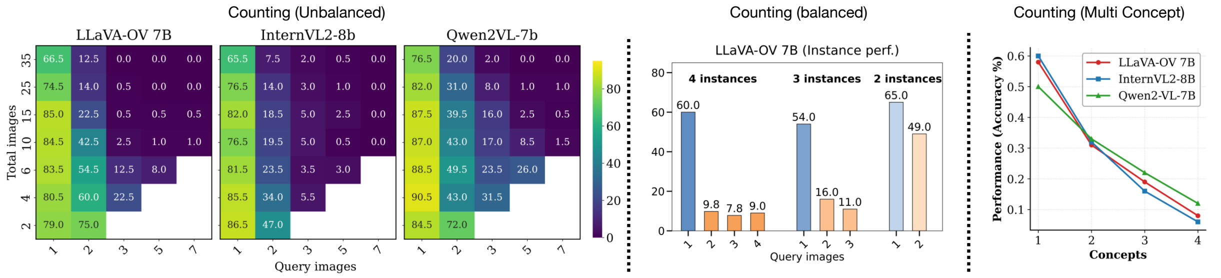

Figure 1: **Counting performance under different settings. Left (Unbalanced):** We compare different LVLMs by analyzing the trade-off between the number of query images and the total number of images without controlling for number of instances. **Mid (Balanced)**: We fix the total number of images to 7 and of object instances distributed across query images to 4,3 and 2. In both settings, performance consistently drops when instances are spread over multiple images. **Right (Multi Concept)**: We increase the complexity by adding more classes (concepts) to the counting task, and observe a steep performance drop, indicating limited capacity for multi-concept tracking.

models can find certain concepts within a long sequence of images. Related works also study more specialized settings, including temporal reasoning in Mementos (Wang et al., 2024d), multi-context visual grounding in MC-Bench (Xu et al., 2025), and egocentric spatial reasoning across frames in (Ravi et al., 2025). In contrast, we provide a more granular analysis of model performance across various controlled dimensions, such as information distribution, query complexity and distractor presence. Moreover, while prior works often repurposed existing datasets, we design our task from scratch to allow for selective performance exploration. This allows us to pinpoint specific strengths and weaknesses in current models that prior benchmarks may have overlooked and, importantly, provide deeper actionable insights on their underlying root causes.

Analysis of LVLMs: Parallel to the development of benchmarks, there is growing interest in analyzing the internal mechanisms of LVLMs to better root-cause their limitations at a data and architecture level. Current studies have investigated issues such as hallucination (Liu et al., 2023; Ouali et al., 2024; Guan et al., 2024; Nath et al., 2026), modality bias (Deng et al., 2025; Wang et al., 2025b), and sensitivity to input phrasing (Qian et al., 2024). These works often involve probing the models with carefully-designed inputs to reveal their decision-making process. Only recently have such analyses expanded to multi-image LVLMs (Wang et al., 2024c; Wu et al., 2025; Sharma et al., 2024). The closest to our work is the study by Wu et al. (Wu et al., 2025), which examines the retrieval capabilities of multi-image LVLMs as the sequence length increases, showing limitations when operating over long sequences. However, their focus is primarily on the models' ability to locate specific items within an image set and does not control for confounding factors or seek to identify the root causes beyond data scarcity.

Instead, we systematically probe additional dimensions of multi-image understanding, such as information aggregation and multi-concept tracking. To control for confounding factors, our evaluation is designed to isolate specific individual aspects of multi-image understanding, leading to precise conclusions and to the identification of areas for improvement. Moreover, we analyze the model's internal behavior, and complement our analysis with proposed solutions to address the identified challenges at both data and optimization levels.

3 Challenges and Insights in Multi-Image LVLMs

Here, we systematically investigate the current LVLMs limitations in multi-image scenarios across six complementary dimensions: information distribution, query complexity, reasoning patterns, robustness to visual distractors, scalability with the number of images, and multi-concept tracking. For this purpose, we introduce MIMIC, a controlled testbed synthesized from a curated subset of MS-COCO (Lin et al., 2014). Using the manually annotated bounding boxes and labels, MIMIC generates multi-image sequences that allow precise control over information spread, distractor presence, object-instance distributions, and sequence length. This design enables decorrelated, fine-grained analyses of the model's behavior. Beyond these dimensions, our framework examines the models' mechanisms for aggregating and reasoning over distributed visual information. Through this controlled analysis, we aim to isolate the specific limitations and offer actionable insights for the next

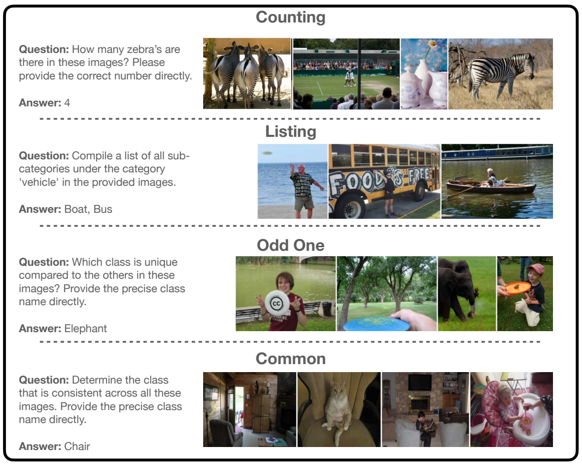

Figure 2: **MIMIC Bench**: examples of each task.

generation of visual understanding models.

Section 3.1 describes the tasks design and dataset construction of our probing benchmark. Section 3.2 describes the systematic probing of several state-of-the-art LVLMs to uncover their strengths and limitations in multi-image understanding.

	Counting		Common	Odd One	Listing	Overall
	Bal.	Unbal.				
Queries	5000	5800	1000	1000	1000	13800
Images	5370	3761	3842	3726	4137	13145
Obj. inst. per query	44.3	132.1	26.1	20.7	27.9	77.0
Min img per query	7	2	3	4	2	2
Max img per query	7	35	8	6	8	35
Median img per query	7.0	15.0	5.0	5.0	5.0	7
Avg words per question	15.1	15.2	14.7	17.3	13.6	15.2

Table 1: MIMIC Benchmark statistics per task. Counting settings: Balanced (Bal) and Unbalanced (Unbal).

3.1 Testbed benchmark construction

We build the probing dataset by procedurally generating multi-image, open-ended question-answering tasks that target distinct aspects of cross-image reasoning. To this end, we sample a curated subset from MS-COCO (Lin et al., 2014) by filtering out images whose object bounding boxes occupying less than 5% of the image in order to ensure visual recognizability at common LVLM input resolutions (e.g. LLaVA-OV's 384×384px). To minimize the impact of potential class imbalance, we first select a pool of classes and then sample from this pool, ensuring that each class is chosen with an equal probability. This ensures that the distributions of classes remain consistent across settings.

MIMIC defines four core tasks: Counting, Listing, Common, and Odd-One, each targeting a distinct facet of multi-image reasoning. Fig. 2 provides some qualitative examples, and Table 1 reports dataset statistics. All tasks are formulated as open-ended question answering rather than multiple-choice to increase challenge, avoid shortcuts introduced by fixed option sets, and remove the need to calibrate distractor choices. To further reduce prompt bias, we employ multiple templatized prompts per task (see appendix for a template list). Below, we describe each task in detail.

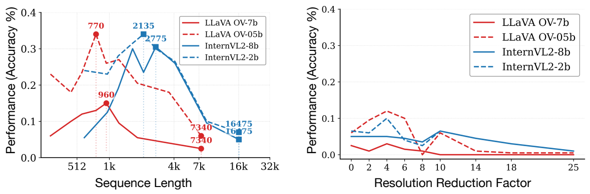

Figure 3: **Effect of vision token sequence length on performance. Left**: Sequence length reduction via 1-D pooling. The square denotes the original sequence length. **Right**: Control experiment reducing the information via pixel space pooling while keeping the sequence length fixed. Results are reported for counting task with 3 query images and total 10 images.

Counting: Given a set of N input images, and a query containing k object classes, the model is asked to count the total number of instances of each class. With increasing difficulty, we vary the distribution of object instances across images. For example, in the easiest setting, all may be concentrated in a single image, while in more challenging cases, instances are spread across multiple images. We refer to this as the *information spread*. Additionally, we introduce distractors - images that do not contain any instances of the target objects - to assess the model's ability to focus on relevant information. In summary, this task offers the following controllable dimensions: (1) number of object classes to count (k); (2) information spread across images (s); (3) number of distractor images and (4) total number of images. Each case probes different aspects and potential biases. For instance, increasing the number of object classes k tests the model's multi-concept tracking ability, while varying the information spread evaluates its capacity to aggregate information across images.

To account for potential biases caused by the long-tail distribution of object instance counts in natural images, which may lead to models favoring smaller counts, we design two distinct settings: (1) *Balanced*, where the total number of object instances is fixed, but distributed across a varying number of images; (2) *Unbalanced*, where the total number of object instances varies arbitrarily with the number of images. The metric of choice is binary accuracy, i.e. the answer is correct if it matches the ground truth count exactly.

Listing: The model is presented with a set of N images and asked to list all object classes belonging to a given category (e.g.: animals, vehicles, etc.) that it can identify. This task evaluates the model's ability to exhaustively extract information in a dense manner from multiple images. As a byproduct, it also measures the model's visual perception ability to recognize and categorize multiple objects, as well as its capacity to aggregate this information into a coherent list. Similar to the Counting task, we vary the number of images and the distribution of object instances to assess the model's robustness in multi-image understanding. The model's response is evaluated on the completeness and accuracy of the list, using the F1 score as metric. See the appendix for a complete hierarchy of object categories and subcategories.

Common and Odd-One: These two tasks are designed to assess the model's ability to identify shared or unique elements across multiple images. Importantly, while previous tasks focus on aggregating information, these tasks require comparative reasoning across images, hence the model must first implicitly identify all objects before performing cross-image analysis. In the Common task, the model has to determine which object class is present in all provided images, while in the Odd-One case, it must identify the object class that is present in a minority of images. For simplicity, we ensure by design that the answers are unique. The model's answers are evaluated based on their correctness, with binary accuracy as the metric.

3.2 Empirical analysis

Setup: We evaluate several state-of-the-art LVLMs: LLaVA-OV (Li et al., 2025), Qwen2-VL (Wang et al., 2024a) and InternVL2 (Chen et al., 2024b). We use publicly available checkpoints and follow the official data processing pipeline. For test data, we use the MIMIC benchmark described in Section 3.1, selecting tasks and configurations that best isolate the dimensions we aim to probe.

Performance in Relation to Sequence Length and Number of Images: LLMs are known to exhibit position and sequence-length biases (Ravaut et al., 2024), with tokens appearing earlier and late in the sequence receiving more attention. Unlike for LLMs, we distinguish two axes of sequence length growth: (1) increasing the number of images, and (2) increasing the input image(s) resolution. We seek to understand if the performance degradation stems from the model's inability to handle long sequences, or from the inability to process many images. We disentangle these two factors with the following experiments: (a) directly increasing the number of images without explicitly controlling for sequence length. In this setting, we simply vary the number of images provided to the model, measuring performance on the counting task. As the results in Fig. 1 (left) show, performance degrades consistently across all settings for all models as the total number of images increases from 2 to 35. (b) reducing the vision token sequence length through 1-D average pooling applied to the original multi-image vision tokens. To ensure that the observed behavior is not an artifact of reduced information, we also perform a control experiment where we similarly reduce the amount of information by downsampling and then upsampling back the images in pixel space, prior to being passed to the vision encoder. This preserves the initial sequence length but reduces the amount of visual information available to the model. This allows us to assess if the performance degradation observed in (a) is primarily due to the increased sequence length or number of images.

The results are summarized in Fig. 3. On the left, we plot performance changes for different models as we reduce the number of vision tokens via 1-D pooling. Due to different processing, each model allocates a different number of tokens per image, hence we mark two points: the extreme right (no downsampling) and the central point that maximizes performance. On the right, we show the control experiment that decreases the information in the pixel space artificially without reducing sequence length. Surprisingly, we find that reducing the sequence length in a zero-shot manner via 1-D pooling up to $4-8\times$ leads to significant performance improvements across all models. The control experiment confirms that gains are due to sequence length reduction rather than *information reduction*. This suggests that the models primarily struggle with long sequence understanding rather than with processing multiple distinct images.

Finding 1: The performance degradation in multi-image scenarios stems primarily from **increased sequence length** rather than the increased number of images.

Moreover, we observe that for LLaVA-OV per-

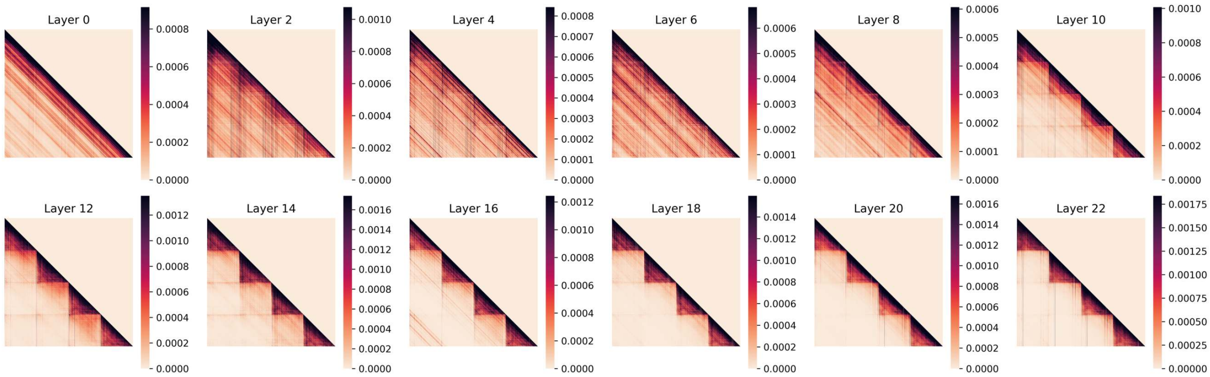

Figure 4: **Inter-image and intra-image token attention across layers.** The attention patterns transitions from cross-image to intra-image interactions as we advance in depth.

formance peaks when the vision sequence length is approximately that of one or two images (i.e., roughly the number of vision tokens for a 384×384 image/patch). This suggests the model effectively relies on a single-image context and has limited practical multi-image integration; we later evaluate how targeted finetuning can mitigate this limitation.

Finding 2: Current LVLMs primarily **behave as single-image models**: performance peaks when the vision-token sequence length matches that produced by one or two images.

Information aggregation across images: Prior benchmarks rarely control for how information is distributed across images, making it difficult to isolate whether models can effectively aggregate information across images. To this end, we vary the *information spread* in the counting task, which defines how object instances are distributed across images. In Fig. 1 (left and middle) we show the results of increasing the number of images containing the object instance from 1 to 7. We observe a sharp drop in accuracy that approaches 0 even when very few distractors are present. This trend is consistent across all models tested and manifests both in balanced and unbalanced counting settings. This indicates that the models may rely on shortcuts, such as focusing on a single or very small subset of images, rather than effectively integrating information from all provided images.

Finding 3: Current LVLMs **struggle to aggregate information** across multiple images.

Robustness to visual distractors. In real-world scenarios, models often encounter irrelevant or distracting information. To evaluate the robustness of LVLMs, we introduce a varying number of irrelevant images into the input sequence. As shown in Fig. 1 (left), the accuracy decreases as the number of distractors increases (e.g: from 79.0% to 66.5% (1 vs 34 distractors) for 1 query image, from 75.0% to 12.5% for two query images, etc., for LLaVA-OV). A particularly pronounced drop occurs as the number of images containing the object of interest increases, suggesting that distractors exacerbate the models' existing difficulties in aggregating information across multiple images.

Finding 4: Models **are sensitive to visual distractors**, especially when information is spread out.

Multi-concept tracking. The ability to track and attend to multiple concepts simultaneously is critical for multi-image understanding. To probe this capability, we vary the number of object classes k that the model is required to count. As shown in Fig. 1 (right), the model performance degrades sharply as k increases, indicating a limited capacity to handle multiple concepts at once.

Finding 5: LVLMs demonstrate **limited capacity for multi-concept tracking**, reducing their reliability on complex multi-object queries.

Multi-image interaction. To probe how visual information is propagated and integrated across images at the token level, we analyze attention patterns among vision tokens in multi-image inputs. Concretely, we compute the normalized attention scores from each vision token to all other vision tokens in the input sequence subject to an autoregressive attention masking on a subset of 50 samples. Fig. 4 summarizes the results across a

Model	Geographic.	Counting	Action.	Grounding	Matching.	Ordering	Scene.	Difference.	Cartoon.	Diagram	Attribute.	Retrieval	Overall
Random Choice	25.0	21.0	23.4	25.0	24.1	22.8	25.0	23.2	25.0	29.6	20.0	21.3	24.0
Human	98.0	94.9	97.6	85.7	94.8	87.5	94.6	92.9	82.1	98.99	87.6	86.3	93.2
GPT-4o (OpenAI, 2023)	56.0	49.2	44.5	36.9	86.9	23.44	71.5	60.3	51.3	88.7	56.1	80.1	68.0
Gemini Pro (Team et al., 2023)	48.0	28.6	36.0	28.6	66.6	12.5	59.1	45.3	47.4	64.8	41.3	43.8	49.4
Mantis-8b-Idefics2 (Jiang et al., 2024b)	26.0	38.5	33.5	26.2	53.9	18.8	57.0	28.8	38.5	67.6	48.5	35.6	44.5
Idefics-9B-Instruct (Laurençon et al., 2023)	35.0	21.8	26.2	26.2	24.8	15.6	56.5	27.6	39.7	25.4	17.9	17.1	35.4
Emu2-Chat(37B) (Sun et al., 2024)	34.0	31.2	27.4	26.2	37.3	15.6	48.4	32.6	43.6	37.7	31.6	24.0	33.6
VILA1.5-13B (Lin et al., 2024)	31.0	19.7	28.7	25.0	40.9	10.9	56.5	24.7	30.8	42.7	24.5	30.1	33.1
LLaVA-NeXT-34B (Liu et al., 2024b)	12.0	36.3	26.2	33.3	37.9	21.9	54.3	22.1	41.0	38.2	38.3	25.0	33.3
LLaVA-v1.5-7B (Liu et al., 2024a)	20.0	23.1	27.4	14.3	23.5	23.4	35.0	20.0	24.4	25.1	23.0	19.9	23.5
LLaVA-v1.5-13B (Liu et al., 2024a)	20.0	25.2	29.3	14.3	20.3	20.3	36.6	20.0	25.6	31.7	23.0	20.9	24.4
CogVLM (Wang et al., 2024b)	13.0	14.1	26.2	16.7	21.3	12.5	41.4	19.7	41.0	19.6	16.3	15.8	20.9
MiniGPT-4-v2 (Chen et al., 2023)	13.0	12.0	14.0	25.0	17.0	18.8	14.5	20.0	21.8	21.6	17.4	14.7	17.4
Qwen2-VL-7B (Wang et al., 2024a)	12.0	38.9	42.7	28.5	57.5	10.9	75.3	32.9	38.5	49.2	46.4	26.7	43.0
Qwen2-VL-2B (Wang et al., 2024a)	14.0	27.8	35.4	26.2	34.3	10.9	51.1	19.4	39.7	21.4	31.1	15.4	27.2
InternVL2-8B (Chen et al., 2024a)	17.0	30.3	34.7	28.5	43.5	17.2	60.2	26.2	46.2	46.5	42.8	33.6	37.9
InternVL2-2B (Chen et al., 2024a)	17.0	21.8	26.8	26.2	31.7	10.9	52.2	17.4	35.9	21.6	16.8	13.7	24.3
LLaVA-OV-0.5B (Li et al., 2025)	22.0	20.9	31.1	25.0	30.2	7.8	46.7	24.1	42.3	25.1	23.9	20.5	26.8
Ours	43.0	20.5	41.4	22.6	38.4	9.4	52.7	22.1	37.2	39.4	29.6	32.5	33.6
Ours (Masked)	31.0	17.5	40.8	33.3	37.3	14.1	53.8	21.5	42.3	34.9	28.1	32.9	32.5
LLaVA-OV-7B (Li et al., 2025)	43.3	24.8	36.6	29.7	45.3	17.2	71.5	30.0	35.9	54.2	32.7	46.2	41.7
Ours (Masked)	44.0	35.9	51.2	42.9	59.9	12.5	71.0	43.5	38.5	62.1	52.0	48.6	51.3

Table 2: Performance comparison across different MuirBench (Wang et al., 2025a) subtasks. **Ours (Masked)**: Our efficient model trained with LoRA and masked attention. **Ours**: Fully fine-tuned model. *Due to computational constraints, we do not fully finetune LLaVA-7B model.*

few layers of interest for a LLaVA-OV model for multi-image inputs with 4 images. We find that in earlier layers, there is a significant amount of inter-image attention, indicating that the model is attempting to integrate information across images. However, as we progress to deeper layers, the attention becomes predominantly intra-image. This inflection point occurs somewhere around the middle of the network. This shift may contribute to the observed difficulty in aggregating information across multiple images. Conceptually, the build-up of representations appears to proceed from broad semantic correlations across images to finer-grained, instance-level integration.

This has a series of consequences: (1) the early inter-image attention may introduce noise or distractions that hinder the model's ability to focus on relevant information in later layers; hence, early mistakes in cross-image interaction are harder to correct; (2) the cross-image interaction under a causal attention mechanism may lead to error propagation, where tokens belonging to later images accumulate a higher amount of noise with incorrect information from earlier images; this may reduce the vision perception capability of the model for later images and explain some of the performance degradation as the number of images increases; (3) the architecture and training objectives may not sufficiently encourage cross-image integration, leading to a default behavior of treating images independently; (4) the observed attention patterns may reflect inherent biases in the training data, where the multi-image tasks don't require deep cross-image reasoning, leading the model to learn shortcuts that prioritize single-image understanding.

Finding 6: Inter-image attention diminishes in deeper layers of LVLMs, indicating a shift from cross-image integration to intra-image focus.

4 Method

In the previous section, we identified key limitations of LVLMs on multi-image tasks via zero-shot evaluation using the MIMIC benchmark. Here, we investigate targeted fine-tuning strategies derived from our findings and aimed at improving multi-image reasoning capabilities. In particular, we explore two complementary approaches: a data-centric fine-tuning strategy using synthetically generated multi-image data, and an optimization-centric attention-masking strategy.

Multi-Image Finetuning: We fine-tune LLaVA-OV models on a unified training dataset composed of samples procedurally generated using the MIMIC pipeline (see section 3.1) together with the original LLaVA-OV multi-image instruction-tuning data (approximately 580K samples). Unlike the MIMIC benchmark used for evaluation, our fine-tuning data is built from OpenImages and provides explicit supervision for cross-image reasoning. It contains approximately 198K samples, with sequence lengths of up to 10 images (see appendix), deliberately exposing models to substantially longer vision-token sequences. All four MIMIC tasks are included to encourage diverse multi-image reasoning behaviors.

Model	MuirBench	Blink	MMIU	MIRB	MMT (val)	NLVR2	Avg.
GPT-4V	62.3	54.6	-	53.1	64.3	-	-
InternVL2-Llama3-76B	51.2	56.8	44.2	58.2	67.4	-	-
LLaVA-v1.5-7B	20.0	37.1	19.2	28.5	-	-	-
InternVL2-2B	24.3	16.3	13.6	25.0	46.7	18.9	24.1
InternVL2-8B	37.9	23.4	36.8	48.6	57.9	8.7	35.6
Qwen2VL-2B	27.2	12.7	38.7	45.9	51.9	41.6	36.3
Qwen2VL-7B	43.0	17.7	52.6	60.8	61.7	41.5	46.2
LLaVA-OV-0.5B	26.8	40.4	34.2	31.8	41.1	61.2	39.3
Ours	33.6	38.9	37.2	32.8	45.6	68.0	42.7
Ours (masked)	32.5	39.1	36.3	28.5	45.9	65.1	41.2
LLaVA-OV-7B	41.7	50.4	45.0	47.2	56.6	84.2	54.2
Ours (masked)	51.3	51.9	45.5	51.0	55.3	87.3	57.1

Table 3: Comparisons on multi-image benchmarks: MuirBench, Blink, MMIU, MIRB, MMT, and NLVR2.

Model	Common	Counting	Odd-one	Listing	Avg.
Mantis-8B-llama3 (Jiang et al., 2024b)	13.0	19.9	10.9	17.0	15.2
InternVL2-2B (Chen et al., 2024a)	25.6	11.7	9.6	19.6	16.6
InternVL2-8B (Chen et al., 2024a)	45.2	18.9	30.2	29.8	31.0
Qwen2VL-2B (Wang et al., 2024a)	41.9	21.7	30.2	23.8	29.4
Qwen2VL-7B (Wang et al., 2024a)	58.6	35.7	35.9	23.4	38.4
LLaVA-OV-0.5B (Li et al., 2025)	44.7	29.7	8.3	22.8	26.4
Ours	68.5	37.8	41.0	34.5	45.5
Ours (masked)	68.9	35.8	50.9	42.0	49.4
LLaVA-OV-7B (Li et al., 2025)	71.5	29.7	58.1	56.6	54.0
Ours (masked)	75.5	51.2	72.1	55.0	63.8

Table 4: Comparisons on our benchmark. We report model's accuracy for Odd-one, Common and Counting whereas f1 score for listing benchmark.

Attention Masking: Our analysis shows that inter-image attention diminishes in deeper layers (see fig. 4). Motivated by this, we apply layer-wise attention masking during fine-tuning, restricting vision tokens to attend only to tokens from the same image in selected layers, while leaving text-token attention unchanged. This design offers two key benefits. First, it reduces unnecessary cross-image interactions, leading to a more efficient model with lower computational cost (see table 9 and fig. 8 of appendix). Second, it encourages cleaner image-local representations in deeper layers, which empirically improves performance across several benchmarks. For this setting, we employ LoRA-based fine-tuning to further improve parameter efficiency. See appendix for implementation details.

5 Results

5.1 Comparison with state-of-the-art

Existing multi-image benchmarks. We first report results on MuirBench (Wang et al., 2025a) and its subtasks in table 2. Across all model sizes, our approach consistently outperforms the corresponding LLaVA-OV baseline. Notably, for the 7B model, our masked-attention variant improves the overall score from 41.7 to 51.3%. We observe a similar trend for the smaller 0.5B variant, indicating that the improvements are robust across model sizes. Interestingly, our method generalizes well to out-of-domain subtasks, including geographic, action and diagram understanding, suggesting that our data construction strategy teaches the model multi-image processing *concepts* rather than *object perception*, which we argue develops already in the single-image training phase.

Next, we extend the evaluation to additional multi-image benchmarks, including Blink (Fu et al., 2024b), MMIU (Meng et al., 2025), MIRB (Zhao et al., 2024), MMT (Ying et al., 2024), and NLVR2 (Suhr et al., 2019). Our approach achieves consistent improvements across all variants. As shown in Tab. 5, our masked-attention fine-tuning strategy yields significant gains over the baseline even with very few trainable parameters, and in some cases outperforms full fine-tuning (e.g., LLaVA-OV 0.5B).

MIMIC benchmark. We report results in Tab. 4. Unless otherwise specified, all results for Counting subtask correspond to the balanced split. Our method significantly outperforms LLaVA-OV across all four tasks. For the 0.5B model, the average score improves from 26.4 to 49.4, while for the 7B model, masked fine-tuning increases performance from 54.0 to 63.8. Gains are most pronounced on the Common and Odd-One tasks, highlighting improved information aggregation and multi-concept reasoning across images.

5.2 Ablation studies and analysis

Cross-task generalization. In this experiment, we train models on individual subtasks (e.g., Counting, Common, Odd-One, and Listing) to analyze their complementary roles and assess cross-task generalization. Table 5 (left) shows the results. We observe that training on the Common task generalizes well to Counting and Listing, but not to Odd-One; a similar trend is observed when training on Odd-One. This behavior is expected, as the two tasks are complementary in nature: Common requires aggregating information across multiple images, whereas Odd-One emphasizes localizing distinctive evidence within a single image. Training on Listing consistently improves performance across all other tasks, while training on Counting primarily benefits Odd-One.

Efficiency analysis. Table 5 (mid) shows that our masked attention variant achieves superior performance with substantially lower computational cost compared with vanilla attention. On the 0.5B back-

	Common	Count	Odd-one	List
Ours (all)	68.5	37.8	41.0	34.5
LLaVA-OV-0.5B	44.7	29.7	8.3	22.8
only Common	73.7	32.0	3.7	30.7
only Counting	35.8	39.4	12.2	20.7
only Odd One	34.4	31.8	53.6	31.1
only Listing	46.0	29.3	11.1	28.3

	FLOPs (Gain)	Avg. Perf.
LLaVA-OV-0.5B	58B (0%)	26.4
Ours	58B (0%)	45.5
Ours (masked)	11.2B (81%)	49.4

Layers masked	Comm.	Count.	Odd.	List.	Avg.
No mask.	70.0	32.0	37.9	44.5	46.1
0-23	64.5	36.1	20.9	29.2	37.7
0-11	62.5	27.3	28.8	33.6	38.1
12-23	68.9	35.8	50.9	42.0	49.4

Table 5: **Left:** Cross-task generalisation. We train LLaVA-OV model individually with each task and compare cross task generalisation. Ours (with all task): upperbound trained with all 4 tasks. **Middle:** Efficiency Analysis of Masked Attention. **Right:** Ablation wrt different layers for attention masking. Experiments on LLaVA-OV 0.5B.

bone, masked finetuning reduces the FLOPs by ∼81%, while outperforming full finetuning. This confirms that selectively constraining inter-image attention is both effective and efficient. See appendix for details on FLOPs estimations.

Attention masking strategy. Table 5 (right) ablates the layers at which attention masking is applied. Masking only deeper layers (layers 12–23) yields the best performance, whereas masking early layers significantly degrades accuracy. These results suggest that early layers are important for effective cross-image information aggregation.

Qualitative analysis. Figure 5 visualizes answer-to-image attention for a 'Counting' example. The baseline fails to attend to relevant objects in the third image, resulting in an incorrect count. In contrast, our model exhibits balanced and semantically grounded attention across all images, leading to the correct prediction. This qualitative evidence corroborates our quantitative improvements.

Figure 5: **Answer-to-Image Attention:** The baseline LLaVA OV (top row) fails to attend to the potted plant in the third image, whereas our method (bottom row) correctly focuses on the relevant object. Visualization is shown at the 15th layer of the LLM.

6 Conclusions

We systematically investigated the capabilities of LVLMs in multi-image contexts through MIMIC, a novel benchmark designed to isolate specific unitary behaviors. Our analysis reveals that current SOTA models fundamentally exhibit "single-image behavior," struggling to aggregate information across inputs or track multiple concepts in the presence of visual distractors. To address this, we introduced a data-centric synthetic fine-tuning strategy and an optimization-centric attention-masking mechanism. These contributions not only address key failure modes but also establish new state-of-the-art results, offering a robust foundation for future research in multi-image understanding.

7 Limitations

While our work offers a rigorous analysis and effective solutions for multi-image LVLMs, we note the following limitations of our study:

- Benchmark Domain: We constructed MIMIC using MS-COCO to maintain precise control over confounding variables (e.g., object counts, occlusion levels). While this design enables exact "unit testing" of model reasoning, extending this controlled methodology to specialized domains, such as dense documents or medical imaging, remains an exciting avenue for future research.

- Resolution Trade-offs: Our analysis demonstrates that reducing sequence length improves multi-image reasoning by mitigating context overload. While highly effective for semantic understanding and counting, tasks requiring pixel-perfect perception of extremely small details might benefit from adaptive resolution strategies, which were outside the scope of this study.

- Architectural Scope: Our proposed analysis focuses on models with open weights. While we expect conclusions to hold for closed models, additional validations (which induce budget constraints) may be useful for reinforcing our findings.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, and 1 others. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: A visual language model for few-shot learning. *Advances in Neural Information Processing Systems*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023. Minigpt-v2: Large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, and 1 others. 2024a. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36:49250–49267.

Ailin Deng, Tri Cao, Zhirui Chen, and Bryan Hooi. 2025. Words or vision: Do vision-language models have blind faith in text? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, and 1 others. 2023. Palm-e: An embodied multimodal language model. *Proceedings of the International Conference on Machine Learning*.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, and 1 others. 2025. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Chaoyou Fu, Yi-Fan Zhang, Shukang Yin, Bo Li, Xinyu Fang, Sirui Zhao, Haodong Duan, Xing Sun, Ziwei Liu, Liang Wang, and 1 others. 2024a. Mme-survey: A comprehensive survey on evaluation of multimodal llms. *arXiv preprint arXiv:2411.15296*.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. 2024b. Blink: Multimodal large language models can see but not perceive. In *Proceedings of the European Conference on Computer Vision*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, and 1 others. 2024. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Wenyi Hong, Weihan Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, and 1 others. 2024. Cogvlm2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, and 1 others. 2024a. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. 2024b. Mantis: Interleaved multi-image instruction tuning. *Transactions on Machine Learning Research*.

Prannay Kaul, Zhizhong Li, Hao Yang, Yonatan Dukler, Ashwin Swaminathan, CJ Taylor, and Stefano Soatto. 2024. Throne: An object-based hallucination benchmark for the free-form generations of large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *Proceedings of the European Conference on Computer Vision*.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, and 1 others. 2020. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal on Computer Vision*, 128(7):1956–1981.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, and 1 others. 2023. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems*.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and 1 others. 2025. Llava-onevision: Easy visual task transfer. *Transactions on Machine Learning Research*.

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

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. 2024. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023. Mitigating hallucination in large multi-modal models via robust instruction tuning. *arXiv preprint arXiv:2306.14565*.

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

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024b. Llava-next: Improved reasoning, ocr, and world knowledge.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2024c. Mmbench: Is your multi-modal model an all-around player? In *Proceedings of the European Conference on Computer Vision*. Springer.

Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 2263–2279.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*.

Fanqing Meng, Jin Wang, Chuanhao Li, Quanfeng Lu, Hao Tian, Jiaqi Liao, Xizhou Zhu, Jifeng Dai, Yu Qiao, Ping Luo, and 1 others. 2025. Mmiu: Multimodal multi-image understanding for evaluating large vision-language models. In *The Thirteenth International Conference on Learning Representations*.

Mriganka Nath, Anurag Das, Jiahao Xie, and Bernt Schiele. 2026. Clipttt: Clip-guided test-time training helps lvlms see better. *arXiv preprint arXiv:2603.26486*.

R OpenAI. 2023. Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2(5):1.

Yassine Ouali, Adrian Bulat, Brais Martinez, and Georgios Tzimiropoulos. 2024. Clip-dpo: Vision-language models as a source of preference for fixing hallucinations in lvlms. In *Proceedings of the European Conference on Computer Vision*.

Yusu Qian, Haotian Zhang, Yinfei Yang, and Zhe Gan. 2024. How easy is it to fool your multimodal LLMs? an empirical analysis on deceptive prompt. In *Neurips Safe Generative AI Workshop 2024*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*.

Mathieu Ravaut, Aixin Sun, Nancy Chen, and Shafiq Joty. 2024. On context utilization in summarization with large language models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 2764–2781.

Sahithya Ravi, Gabriel Herbert Sarch, Vibhav Vineet, Andrew D Wilson, and Balasaravanan Thoravi Kumaravel. 2025. Out of sight, not out of context? egocentric spatial reasoning in vlms across disjoint frames. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 16146–16161.

Aditya Sharma, Michael Saxon, and William Yang Wang. 2024. Losing visual needles in image haystacks: Vision language models are easily distracted in short and long contexts. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Ilias Stogiannidis, Steven McDonagh, and Sotirios A Tsaftaris. 2025. Mind the gap: Benchmarking spatial reasoning in vision-language models. In *Greeks in AI Symposium 2025*.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428.

Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2024. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Fei Wang, Xingyu Fu, James Y Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, and 1 others. 2025a. Muirbench: A comprehensive benchmark for robust multi-image understanding. In *The Thirteenth International Conference on Learning Representations*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Song XiXuan, and 1 others. 2024b. Cogvlm: Visual expert for pretrained language models. *Advances in Neural Information Processing Systems*.

Xidong Wang, Dingjie Song, Shunian Chen, Junyin Chen, Zhenyang Cai, Chen Zhang, Lichao Sun, and Benyou Wang. 2024c. Longllava: Scaling multi-modal llms to 1000 images efficiently via a hybrid architecture. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Xiyao Wang, Yuhang Zhou, Xiaoyu Liu, Hongjin Lu, Yuancheng Xu, Feihong He, Jaehong Yoon, Taixi Lu, Fuxiao Liu, Gedas Bertasius, and 1 others. 2024d. Mementos: A comprehensive benchmark for multi-modal large language model reasoning over image sequences. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 416–442.

Zhaochen Wang, Bryan Hooi, Yiwei Wang, Ming-Hsuan Yang, Zi Huang, and Yujun Cai. 2025b. Text speaks louder than vision: Ascii art reveals textual biases in vision-language models. In *Second Conference on Language Modeling*.

Tsung-Han Wu, Giscard Biamby, Jerome Quenum, Ritwik Gupta, Joseph E Gonzalez, Trevor Darrell, and David M Chan. 2025. Visual haystacks: A vision-centric needle-in-a-haystack benchmark. *International Conference on Learning Representations*.

Yunqiu Xu, Linchao Zhu, and Yi Yang. 2025. Mc-bench: A benchmark for multi-context visual grounding in the era of mllms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17675–17687.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, and 1 others. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, and 1 others. 2024. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. In *International Conference on Machine Learning*, pages 57116–57198.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Bingchen Zhao, Yongshuo Zong, Letian Zhang, and Timothy Hospedales. 2024. Benchmarking multi-image understanding in vision and language models: Perception, knowledge, reasoning, and multi-hop reasoning. *arXiv preprint arXiv:2406.12742*.

Yilun Zhao, Haowei Zhang, Lujing Xie, Tongyan Hu, Guo Gan, Yitao Long, Zhiyuan Hu, Weiyuan Chen, Chuhan Li, Zhijian Xu, and 1 others. 2025. Mmvu: Measuring expert-level multi-discipline video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Kaizhi Zheng, Xuehai He, and Xin Eric Wang. 2024. Minigpt-5: Interleaved vision-and-language generation via generative vokens. *arXiv preprint arXiv:2310.02239*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, and 1 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

A Appendix

A.1 Additional Ablations and Analysis

Multi-image vs single composite image (stitching). To further understand whether the observed failure modes are due to the presence of multiple images or simply due to the increased number of vision tokens, we conduct an additional experiment involving multi-image stitching: Here, we create a single composite image by stitching multiple images together in a grid format, ensuring that the total number of vision tokens remains similar to that of the multi-image input. The conceptual changes that occur here are twofold: (1) the chat template is devoid of any image separators, and (2) the vision encoder may jointly process parts of different images together. The results across all tasks are summarized in Table 6. As the results show, general performance remains similar or increases slightly in some cases.

Ablations wrt Data Components. We conduct an ablation using the 0.5B LLaVA variant of our method to isolate the contribution of MIMIC-generated data relative to the original LLaVA-OV (OneVision) multi-image data. We compare three training setups: (i) OneVision only, (ii) MIMIC only, and (iii) OneVision + MIMIC. As shown in table 7, MIMIC substantially outperforms OneVision on both the MIMIC tasks and the external benchmarks, indicating that the gains are primarily driven by the MIMIC-generated data. Combining MIMIC with OneVision further improves performance on the external benchmarks, suggesting complementary benefits for generalization.

Ablations wrt Attention Masking. We conduct an ablation to assess the effect of attention masking by comparing two training setups: (i) LoRA and (ii) LoRA + Masking. As shown in table 8, adding attention masking improves overall performance on the MIMIC benchmark, particularly on Counting and Odd One, resulting in a higher average score. Importantly, this improvement comes while maintaining comparable performance on external benchmarks, suggesting that masking helps the model attend to relevant cross-image information without harming generalization.

Model	Common	Counting	Odd-one	Listing
LLaVA-OV-0.5B	33.6	28.9	8.8	25.0
LLaVA-OV-0.5B (stitched)	41.0	30.8	22.7	19.5
LLaVA-OV-7B	72.4	29.1	58.0	49.4
LLaVA-OV-7B (stitched)	68.2	35.9	67.1	51.5
Qwen2-VL-2B	40.4	21.8	26.3	50.8
Qwen2-VL-2B (stitched)	36.2	38.8	26.8	51.1
Qwen2-VL-7B	61.3	36.6	35.3	50.3
Qwen2-VL-7B (stitched)	51.7	35.8	51.1	58.5
InternVL2-2B	25.6	28.8	7.1	35.3
InternVL2-2B (stitched)	29.5	29.3	6.2	39.8
InternVL2-8B	42.8	30.9	24.4	52.9
InternVL2-8B (stitched)	42.5	27.3	37.3	54.8

Table 6: Stitching Experiment. To ensure a similar number of tokens for both stitched and multi-image, we resize images to 384×384 for LLaVA-OV and 484×484 for Qwen2-VL and InternVL2.

Efficiency and FLOPs Analysis. Table 9 shows that our masked attention variant achieves superior performance with substantially lower computational cost compared to vanilla attention. On the 0.5B backbone, masked finetuning reduces FLOPs by approximately 81%, while also outperforming full finetuning. This demonstrates that selectively constraining inter-image attention is both effective and computationally efficient.

Formally, given N_t text tokens and N_v visual tokens distributed across M independent images, the FLOPs of a standard transformer layer with full self-attention scale as $\mathcal{O}((N_t+N_v)^2 d + (N_t+N_v)d^2)$, where d denotes the hidden dimension. The first term corresponds to self-attention computation, while the second accounts for the MLP.

In contrast, our masked attention restricts visual tokens to attend only within their respective image blocks, each of size $n = \frac{N_v}{M}$, while preserving global visibility for text tokens. This modifies the complexity to

$$\text{FLOPs} \approx \underbrace{\left(N_t(N_t+N_v) + \sum_{i=1}^{M} n_i^2 \right) d}_{\text{Masked Attention}} + \underbrace{(N_t+N_v)d^2}_{\text{MLP}}. \tag{1}$$

Assuming uniform image sizes ($n_i = \frac{N_v}{M}$), this enables scaling to a larger number of high-resolution images under fixed memory and compute budgets. In our MIMIC benchmark, we observe $M = 10.4$ images per sample with an average of $N_t = 17.4$ text tokens. Since LLaVA-OV uses 730 visual tokens per image, this yields $N_v = 7592$. We use hidden dimensions $d = 896$ for the 0.5B model and $d = 1536$ for the 1.5B variant, consistent with the FLOPs reductions reported in Table 9.

Performance vs. number of images. In Section 3.2 and Fig. 1, we analyze how the perfor-

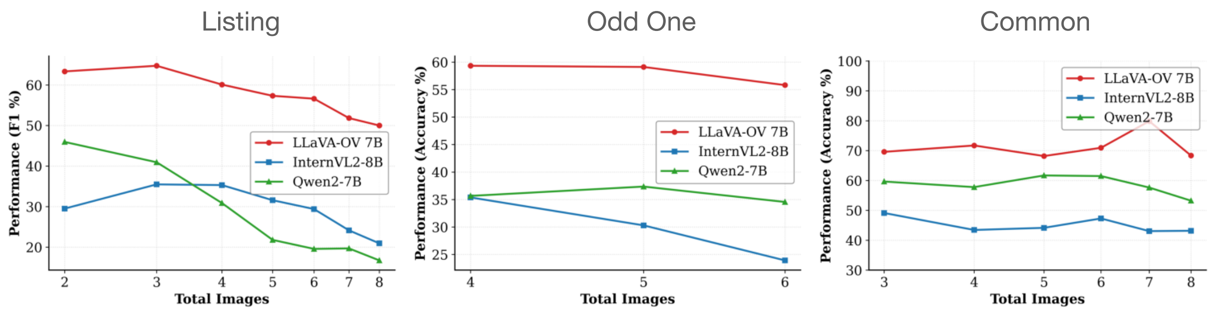

Figure 6: **Performance vs. Number of Images**. We report performance as a function of the total number of input images for the Listing, Odd-One tasks, and common tasks.

	Common	Counting	Odd One	Listing (F1)	Avg
OneVision only	44.7	29.7	8.3	22.8	26.4
MIMIC only	73.6	43.3	64.2	43.8	56.2
OneVision + MIMIC	68.5	37.8	41.0	34.5	45.5

	Blink	MMIU	MuirBench	Average
OneVision only	40.4	34.2	26.8	33.8
MIMIC only	38.9	35.6	33.1	36.0
OneVision + MIMIC	38.9	37.2	33.6	36.6

Table 7: **Performance comparison with different training data components.** MIMIC yields most of the gains, while combining it with OneVision improves external benchmark performance.

	Common	Counting	Odd One	Listing (F1)	Avg
LoRA	70.1	32.0	37.9	44.5	46.1
LoRA + Masking	68.9	35.8	50.9	42.0	49.4

	Blink	MMIU	MuirBench	Average
LoRA	38.0	35.8	34.7	36.2
LoRA + Masking	39.1	36.3	32.5	36.0

Table 8: **Performance comparison with and without masking.** Masking consistently improves performance on the MIMIC benchmark while largely preserving performance on external benchmarks.

	FLOPs (Gain)	Avg. Perf.
LLaVA-OV-0.5B	58B (0%)	26.4
Ours	58B (0%)	45.5
Ours (masked)	11.2B (81%)	49.4
LLaVA-OV-1.5B	107B (0%)	29.8
Ours	107B (0%)	54.7
Ours (masked)	26.7B (75%)	46.4

Table 9: **Efficiency analysis.** Our masked-attention variant is substantially more computationally efficient than the vanilla-attention version used in our method.

Prompt	Accuracy
Determine the class that is consistent across all these images.	43.8
Identify the class that appears in all these images.	47.4
Find the common category in these images.	54.9
Ours (multiple prompts)	44.7

Table 10: Effect of prompt variations on the Common subtask. Performance varies noticeably across individual prompt templates, supporting the use of multiple prompts to reduce prompt bias and improve robustness.

mance of the counting task varies with the total number of images and the number of query images. Here, we extend this analysis by also considering the Listing, Odd-One, and Common tasks. Fig. 6 reports the performance on these three tasks as the number of input images increases. We observe that for the Listing and Odd-One tasks, performance decreases as the number of total images grows, while performance on the Common task remains largely stable and is not affected by the number of images. This behavior is expected, since the Common task contains no distractors and therefore does not require exhaustive reasoning over all images to predict the correct class.

Effect of Prompt Variations. To assess prompt sensitivity, we conduct an additional experiment using individual prompt templates. We evaluate two subtasks, Common and Counting, with three prompt variants each (shown in fig. 12), while keeping the connector fixed across prompts (e.g., "Please provide the precise class name directly" for Common and "Provide the exact number directly" for Counting). Results are shown in tables 10 and 11. We observe noticeable performance variation across individual prompt formulations, indicating that model performance is sensitive to prompt wording even when the task and answer format remain unchanged. These results support the use of multiple prompt templates to reduce prompt bias and improve robustness.

Ablation wrt Masking Layers. We further analyze the sensitivity of our attention masking strategy to the choice of masked layers. Due to computational constraints, we conduct this ablation on the 0.5B model, which also informs our design choice

Prompt	Accuracy
Count the occurrences of <class_id> in these images.	26.7
How many <class_id> are in the images?	31.9
What is the total number of <class_id> present in these images?	30.9
Ours (multiple prompts)	29.7

Table 11: Effect of prompt variations on the Counting subtask. Performance varies across individual prompt templates, while using multiple prompts provides a more robust evaluation setup.

Layers Masked	Layer Range	Common	Counting	Odd One	Listing (F1)	Average
100%	0–23	66.2	37.5	11.8	27.1	35.7
50%	12–23	68.9	35.8	50.9	42.0	49.4
	0–11	66.0	31.5	12.6	34.0	36.0
25%	0–5	71.7	32.6	37.2	41.8	45.8
	6–11	68.4	46.7	39.8	44.1	49.8
	12–17	69.4	47.5	40.9	43.5	50.3
	18–23	70.3	48.4	40.7	44.2	50.9

Table 12: Ablation of attention masking across different layer ranges on the 0.5B model. Later-layer masking consistently outperforms early-layer masking, while masking the last 50% of layers provides a strong trade-off between effectiveness and efficiency.

for larger models. As shown in Table 12, masking later layers is substantially more effective than masking earlier layers, and partial masking outperforms masking all layers. In particular, masking the last 50% of layers provides a strong accuracy-efficiency trade-off.

Extended multi-image interaction. Fig. 4 in Section 3.2 reports the normalized attention scores from each vision token to all other vision tokens in an input sequence of 4 images for a LLaVA-OV model. The scores are computed on a subset of 50 samples and then averaged. Fig. 7 extends this analysis to an input sequence of 6 images. From the figures, we observe that the observations made for 4 images also hold for 6 images. In particular, in the early layers, there is a large amount of inter-image attention, while in deeper layers the attention becomes mostly intra-image, indicating that the model focuses more on individual images. Therefore, we can infer that the observed behavior is intrinsic to the model and does not depend on the number of input images.

Counting (balanced) Performance Comparison. We extend the results from table 4 with a fine-grained analysis of performance improvements in fig. 9. We observe that fine-tuning substantially improves performance when object instances are distributed across multiple query images. For example, when four instances are spread across four images, accuracy increases from 9% to 45.8%.

Model	MuirBench	Blink	MMIU	MIRB	MMT (val)	NLVR2	Avg.
GPT-4V	62.3	54.6	-	53.1	64.3	-	-
InternVL2-Llama3-76B (Chen et al., 2024a)	51.2	56.8	44.2	58.2	67.4	-	-
LLaVA-v1.5-7B	20.0	37.1	19.2	28.5	-	-	-
InternVL2-2B (Chen et al., 2024a)	24.3	16.3	13.6	25.0	46.7	18.9	24.1
InternVL2-8B (Chen et al., 2024a)	37.9	23.4	36.8	48.6	57.9	8.7	35.6
Qwen2VL-2B (Wang et al., 2024a)	27.2	12.7	38.7	45.9	51.9	41.6	36.3
Qwen2VL-7B (Wang et al., 2024a)	43.0	17.7	52.6	60.8	61.7	41.5	46.2
LLaVA-OV-0.5B (Li et al., 2025)	26.8	40.4	34.2	31.8	41.1	61.2	39.3
Ours	33.6	38.9	37.2	32.8	45.6	68.0	42.7
Ours (masked)	32.5	39.1	36.3	28.5	45.9	65.1	41.2
LLaVA-OV-1.5B (Li et al., 2025)	31.1	36.4	33.4	37.7	47.5	70.9	42.8
Ours	39.7	40.0	38.9	36.0	48.8	73.7	46.2
Ours (masked)	32.3	42.2	35.5	30.6	48.1	69.0	43.0
LLaVA-OV-7B (Li et al., 2025)	41.7	50.4	45.0	47.2	56.6	84.2	54.2
Ours (masked)	51.3	51.9	45.5	51.0	55.3	87.3	57.1

Table 13: Extended Comparisons including LLaVA-OV 1.5B with the state-of-the-art on multi-image benchmarks: MuirBench, Blink, MMIU, MIRB, MMT and NLVR2.

Similar gains are observed across different instance distributions, indicating improved cross-image information aggregation.

Extended Performance Comparison with LLaVA-OV 1.5B. We extend the comparison to multi-image benchmarks from table 3 by including the LLaVA-OV 1.5B model. We observe that both of our fine-tuned models outperform the baseline. In particular, our fine-tuned model achieves a 3.4% improvement over the baseline, demonstrating enhanced multi-image reasoning capabilities.

Bigger and latest models. Fig. 10 reports results on the Counting task by increasing the number of images that contain the object instance from 1 to 7. We consider larger (LLaVA-OV 72B) or more recent (Qwen2.5-7B, Qwen3-VL-8B) models compared to those analyzed in Fig. 1 (left). Consistent with previous findings outlined in Section 3.2, even more powerful models achieve strong performance when the object of interest appears in a single query image, but performance decreases as the same object is spread across multiple images.

A.2 Implementation Details.

Training Data. Our method is trained on a unified training set composed of synthetic samples generated using the MIMIC pipeline (see Section 3.1) together with the original LLaVA-OV multi-image instruction-tuning data (approximately 580K samples). Unlike the MIMIC benchmark used for evaluation, the synthetic MIMIC training dataset is built from OpenImages (Kuznetsova et al., 2020) annotations and provides explicit supervision for cross-image reasoning. Additionally, it supports multi-turn conversations and option-based responses (see fig. 11). Dataset statistics are reported in table 14. In particular, the dataset consists of approximately 50K samples from each MIMIC

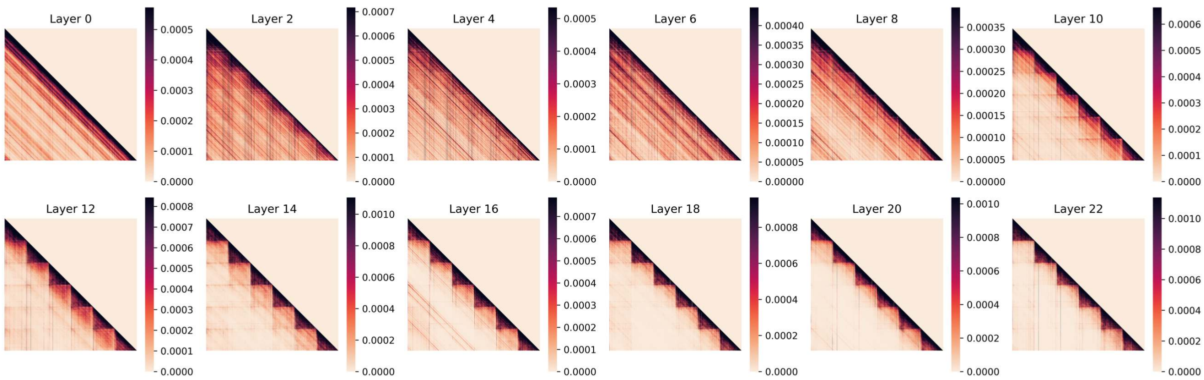

Figure 7: **Inter-image and intra-image token attention across layers for 6 images.**

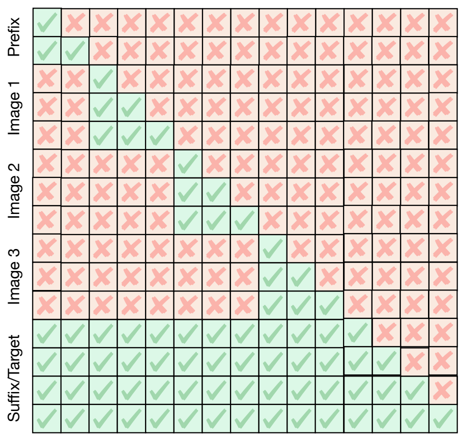

Figure 8: **Our Masked Attention.** Vision tokens are restricted to attend only to tokens from the same image, following a block-diagonal attention pattern, while text tokens in both the prefix and suffix follow the standard autoregressive attention.

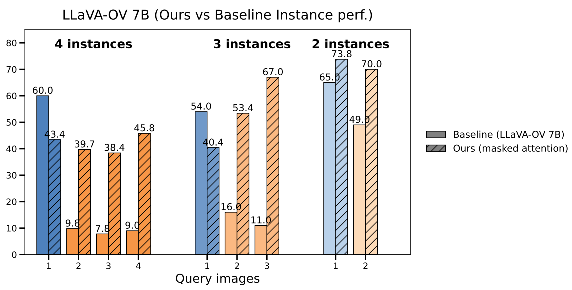

Figure 9: **Comparison on Counting (balanced)**. Our fine-tuned model performs substantially better when object instances are distributed across multiple images. For example, when four instances are spread across four images, performance improves from 9% to 45.8%, indicating enhanced cross-image information aggregation.

subtask, with sequences containing up to 10 images. This design encourages learning effective cross-image information aggregation while retaining general vision-language capabilities through joint training with LLaVA-OV data.

Training Procedure. For both training strategies—the data-centric fine-tuning approach and the optimization-centric attention-masking approach—we start from the LLaVA-OV Single-Image variant (Stage 2.1) (Li et al., 2025), which has been pre-trained on high-quality single-image instruction-tuning data. We freeze the vision encoder and train only the language model layers and the vision-to-language projector.

For the optimization-centric attention-masking strategy, we do not fully fine-tune the language model layers; instead, we apply LoRA adapters to these layers. In addition, we introduce attention masking in the self-attention blocks, restricting vision tokens to attend only to tokens from the same image. For the data-centric fine-tuning strategy, we fully fine-tune the language model layers without attention masking.

All models are trained with an effective batch size of 128, zero weight decay, and a cosine learning rate schedule with a warmup ratio of 0.03. For the data-centric approach, we use a learning rate of $2.5 \times 1e^{-6}$, while for the LoRA-based attention-masking strategy, we increase the learning rate to $2.5 \times 1e^{-5}$. Training is performed on 8 NVIDIA H100 GPUs with approximately 80 GB of memory each. For the attention-masking strategy, we set the LoRA rank to 128.

A.3 Additional MIMIC details

Prompt templates Fig. 12 shows the different prompt templates used for the four tasks. For each task, a prompt is constructed by randomly sampling one template from the task-specific template set (P_{task}) and one from the connector template set ($P_{\text{connector}}$). The two templates are then combined to form the final prompt, i.e., $P = P_{\text{task}} \| P_{\text{connector}}$.

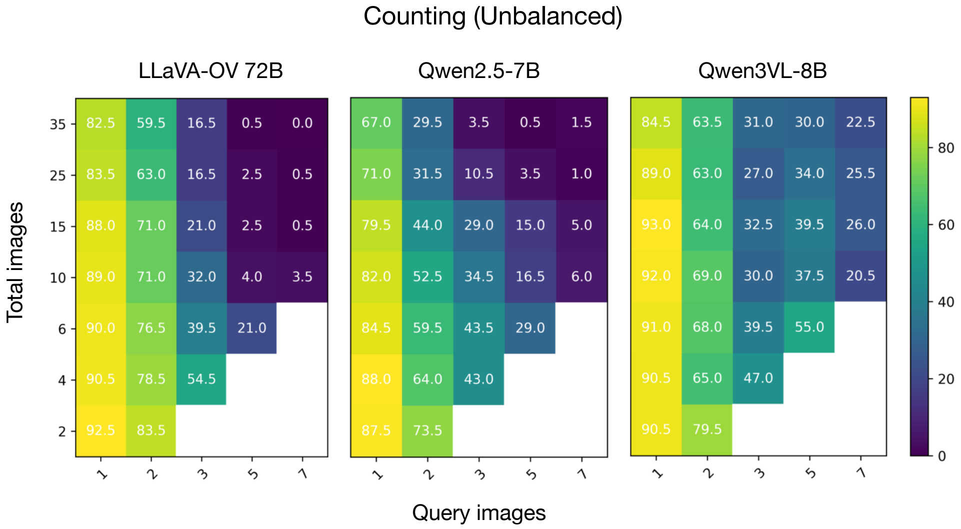

Figure 10: **Counting (unbalanced) performance with bigger and latest models**. We analyze the trade-off between the number of query images and the total number of images for bigger (LLaVA-OV 72B) and latest models (Qwen2.5-7B and Qwen3VL-8B).

Figure 11: **Samples from the MIMIC training dataset**.Unlike the evaluation benchmark, the training data follows the LLaVA data format and includes multi-turn conversations with option-based answers.

	Counting	Common	Odd One	Listing	Overall
Queries	50000	50000	47561	49267	196828
Images	157995	80438	81642	108300	232196
Obj. inst. per query	27.3	16.3	17.8	18.6	20.0
Min img per query	2	2	3	2	2
Max img per query	10	8	8	8	10
Median img per query	5	4	4	4	4
Avg words per question	48.9	44.6	47.7	52.1	48.3

Table 14: MIMIC synthetic training dataset statistics based on OpenImagesv7 (Kuznetsova et al., 2020).

Additional examples We report additional sample instances from the Common, Odd-One, Listing, and Counting categories, in Figs. 13, 14, 15 and 16, respectively.

Task Templates - Common

1. Which class is common in these images?
2. Identify the class that appears in all these images.
3. Find the common category in these images.
4. Determine the class that is consistent across all these images.
5. Which class is repeated in all these images?

Connector Templates

1. Please provide the correct class name directly.
2. Provide the precise class name directly.

Task Templates - Odd One

1. Which class is odd one out in these images?
2. Identify the class that is different from the others in these images.
3. Determine the class that doesn't belong in these images.
4. Find the class that is the outlier among these images.
5. Which class is unique compared to the others in these images?
6. Identify the class that is distinct from the rest in these images.
7. Determine the class that is not like the others in these images."

Connector Templates

1. Please provide the correct class name directly.
2. Provide the precise class name directly.

Task Templates - Listing

1. List all the sub-categories belonging to the category <class_id> in the given images.
2. Provide a list of all sub-categories under the category <class_id> in these images.
3. Enumerate the sub-categories that belong to the category <class_id> in the given images
4. Generate a list of all sub-categories for the category <class_id> in the images.
5. Compile a list of all sub-categories under the category <class_id> in the provided images.
6. Identify and list all sub-categories belonging to the category <class_id> in these images.

Connector Templates

1. Please provide the list of category names directly. Separate multiple names with commas.

Task Templates - Counting

1. How many <class_id> are there in the images?
2. Count the occurrences of <class_id> in these images.
3. What is the total number of <class_id> present in these images?
4. Provide the count of <class_id> visible in these images.
5. How many times does <class_id> appear in these images?
6. How many instances of <class_id> can be identified in these images?

Connector Templates

1. Please state the exact number directly.
2. Provide the exact number directly.
3. Please give the correct number without any extra information.

Figure 12: **Prompt Templates for Different Tasks.** For each task, a prompt is constructed by randomly sampling one template from the task-specific template set (P_{task}) and one from the connector template set ($P_{\text{connector}}$). The two templates are then combined to form the final prompt, i.e., $P = P_{\text{task}} \| P_{\text{connector}}$.

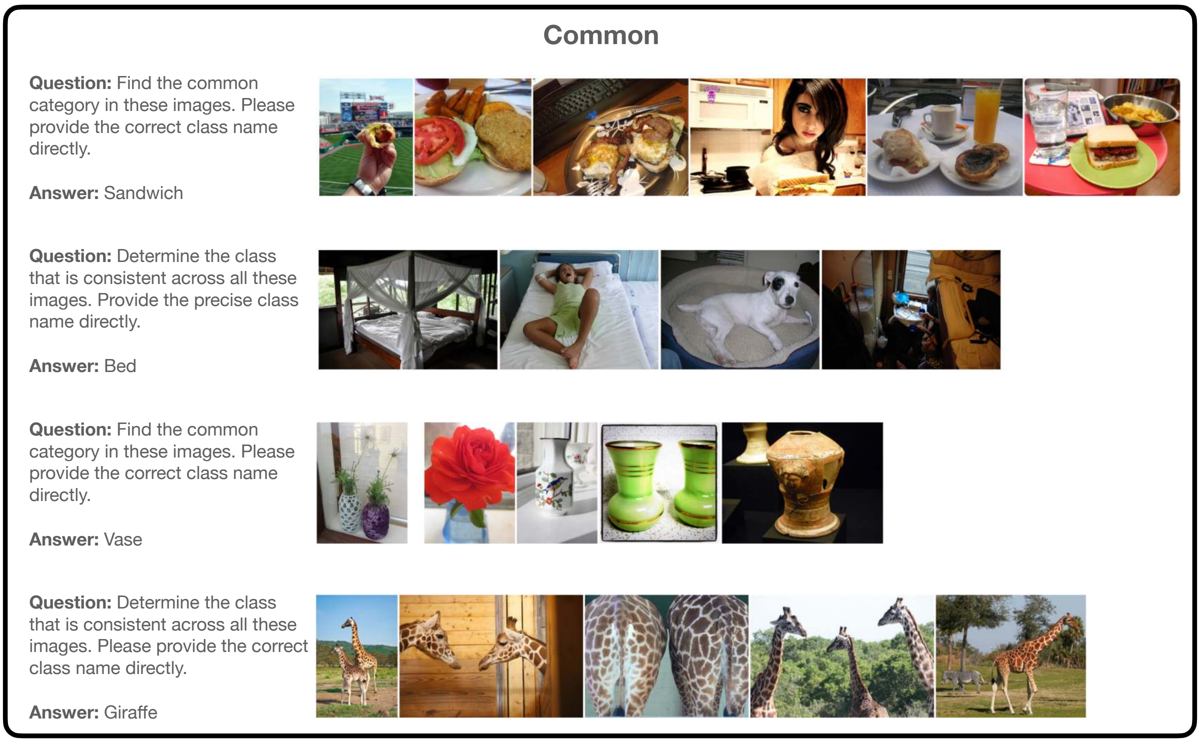

Figure 13: Our MIMIC Evaluation Benchmark. We show some samples from the category 'common' from our benchmark.

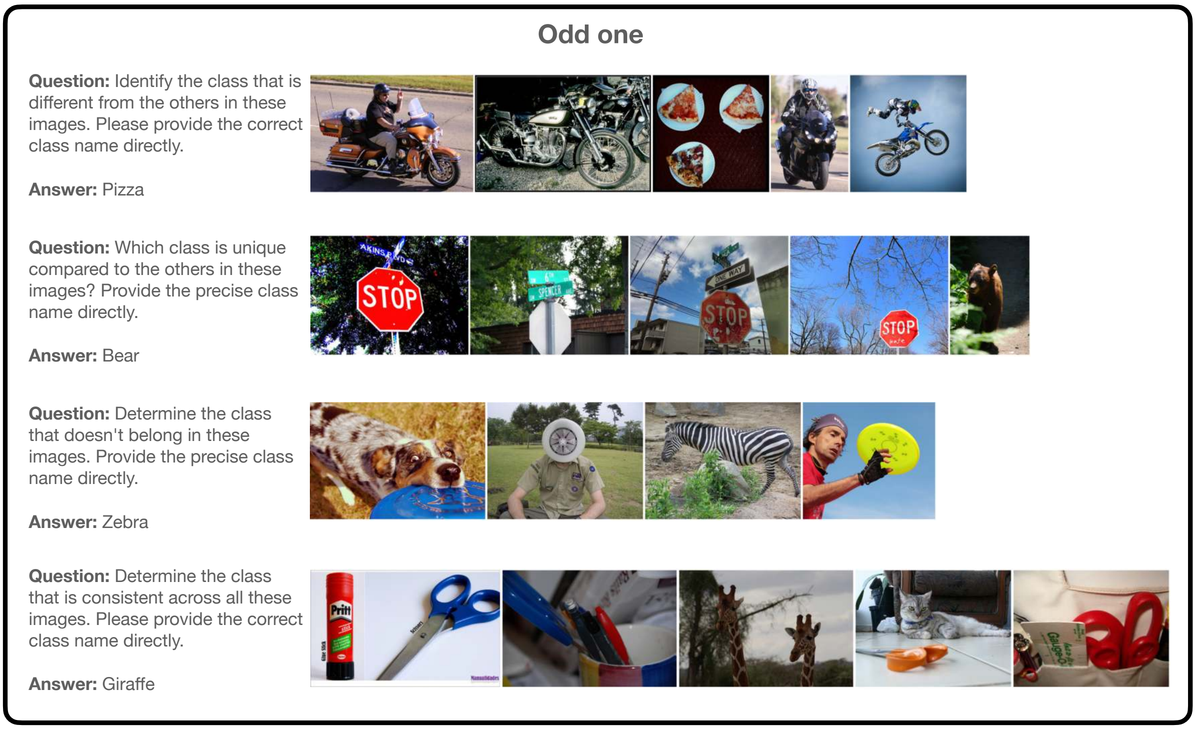

Figure 14: Our MIMIC Evaluation Benchmark. We show some samples from the category 'Odd One' from our benchmark.

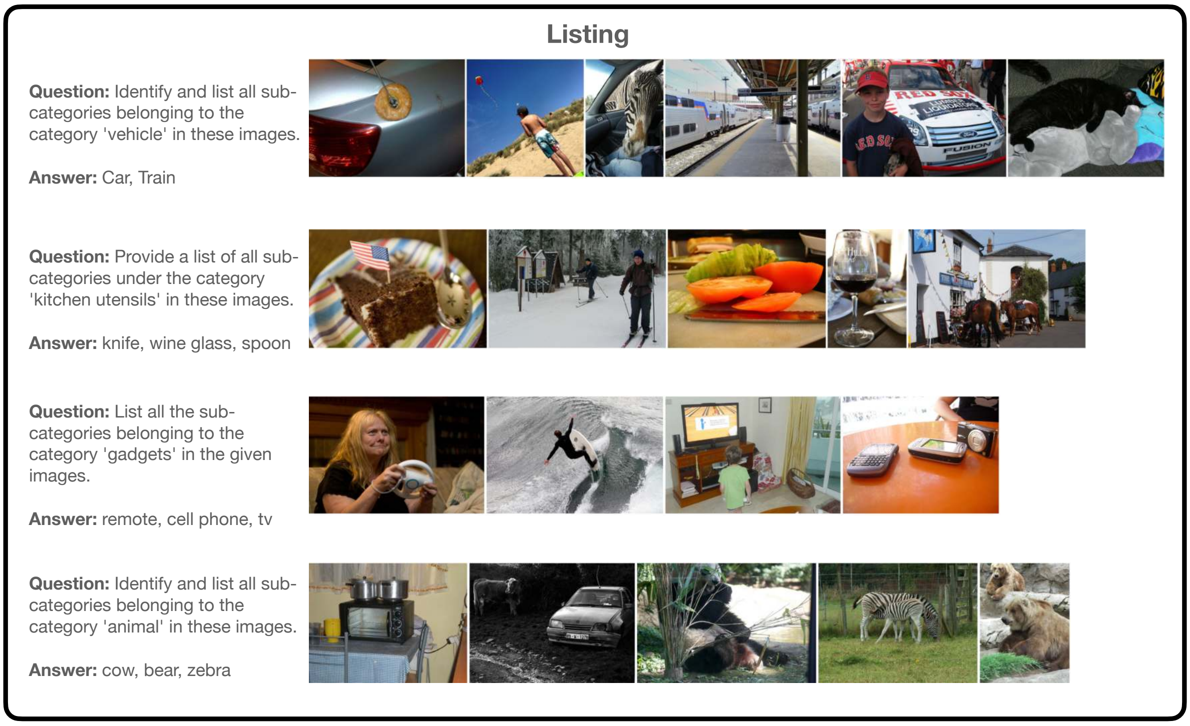

Figure 15: Our MIMIC Evaluation Benchmark. We show some samples from the category 'Listing' from our benchmark.

Figure 16: Our MIMIC Evaluation Benchmark. We show some samples from the category 'Counting' from our benchmark.