StochasT: Learning with Stochastic Turn Depth for Visual Instruction Tuning

Yuan Qing[1], Chengzhi Mao[2], and Boqing Gong[1]

[1] Boston University, Boston, USA
[2] Rutgers University, New Brunswick, USA
{ymqing, bgong}@bu.edu, chengzhi.mao@rutgers.edu

Abstract. Large Vision-Language Models (LVLMs) rely extensively on Visual Instruction Tuning (VIT) to elicit their multimodal reasoning capabilities. However, we find a discrepancy: VIT often packs multiple language tasks about the same image for conversational, multi-turn training, whereas existing benchmarks evaluate LVLMs in isolated, single-turn scenarios. The models can suffer from visual attention decay and contextual overfitting during multi-turn training, making it hard for them to realize their full potential in the mismatched test phase. To close the gap, we propose learning with Stochastic Turn Depth (StochasT), which stochastically groups language tasks for the same image into clusters of varying sizes (turn depth) while preserving their organic order. Hence, while StochasT draws on Dropout and stochastic depth for ResNets, it does not actually drop anything to maximize the utility of the training data. Furthermore, we introduce a challenging, benchmark-agnostic evaluation mechanism based on the Balanced Latin Square to measure LVLMs' robustness under varying contextual dependencies. Extensive experiments demonstrate that StochasT effectively grants LVLMs strong, harmonized capabilities for both single-turn and multi-turn use cases. Code is available at: https://yuanqing-ai.github.io/StochasT.

Keywords: Large Vision-Language Models · Single-Turn, Multi-Turn, and Stochastic Turn-Depth Evaluation · Visual Instruction Tuning

1 Introduction

Large Vision-Language Models (LVLMs) [2, 3, 11, 48, 51, 59] have demonstrated remarkable capabilities across various vision applications, spanning Perception (e.g., object detection, segmentation, OCR), Reasoning (e.g., visual question answering, multi-hop reasoning, chart understanding), and Action (e.g., visual instruction following, embodied navigation, GUI control). Standard approaches for building such systems typically integrate a pretrained Large Language Model (LLM) [19,44] with a pretrained visual encoder, bridged by an alignment module such as an MLP projection layer or a Q-Former [30].

To unlock the multimodal reasoning and generalization potential of LVLMs, Visual Instruction Tuning (VIT) [33,34,71] is critical. Prior studies suggest that much of the world knowledge embedded within foundational LLMs is acquired

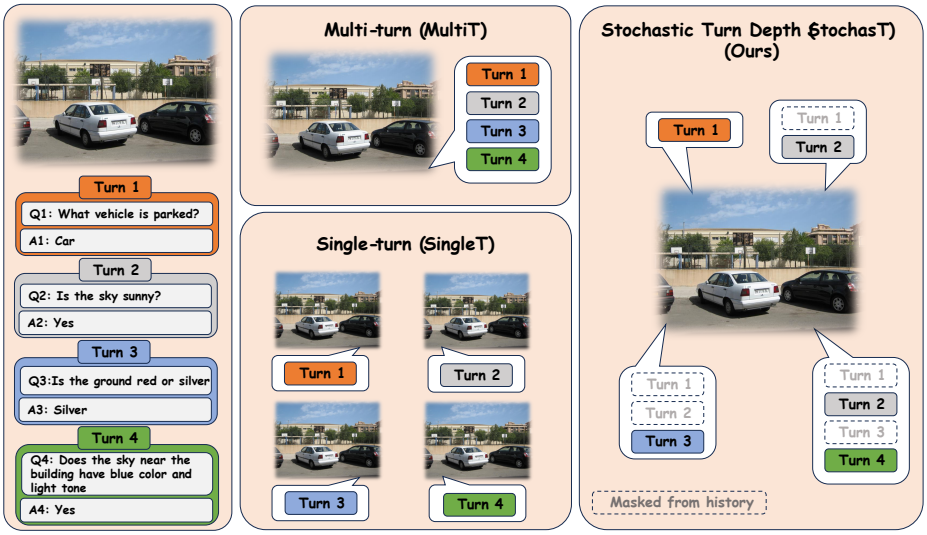

Fig. 1: A visual instruction tuning example (left) and three grouping mechanisms: multi-turn (multiT), singleT, and our proposed stochastic turn depth (**StochasT**).

during large-scale pretraining [50, 69]. Consequently, VIT primarily serves to activate and align this latent knowledge toward downstream multimodal task objectives, rather than learning it from scratch.

Unlike pure language tasks, the high information density inherent to visual data naturally affords multi-turn (multiT) language queries. A single image often grounds multiple distinct instructions (as illustrated in Fig. 1), and this one-image-multiT format has been frequently used in VIT [34]. However, a significant discrepancy exists between this multiT training paradigm and currently prevalent single-turn (singleT) evaluation protocols. As depicted in Fig. 1, multiT training groups all instruction-answer pairs about the same image into one training example, while singleT evaluation anchors every instruction individually to the image. As a result, a model may fail to answer a simple question in isolation but succeed when the exact same question is contextualized within a conversation. Most existing LVLM benchmarks [18, 28, 37, 66] employ singleT testing exclusively, treating related questions about the same image as independent, isolated runs. However, if we

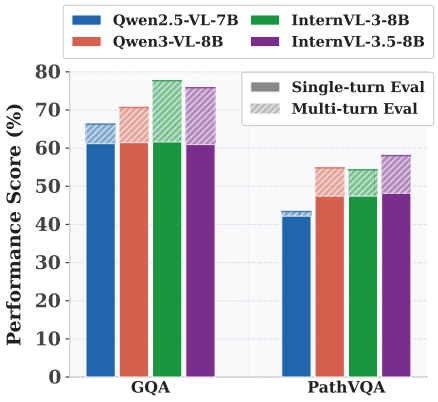

Fig. 2: Performance comparison of several SOTA LVLMs under singleT and multiT evaluation.

instead evaluate LVLMs in the same, multiT way as used in training, we can boost their performance significantly (see Fig. 2; experiment setup in supplement). Surprisingly, the literature has largely overlooked this discrepancy between multiT VIT and singleT evaluation.

In this work, we investigate two primary research questions hinging on this observation: 1) How can we effectively balance the one-image-multiT training paradigm to optimize performance across both singleT and multiT evaluation? 2) Given the substantial discrepancies between singleT and multiT performance, how should we best evaluate LVLMs' robustness in various use cases?

To address the first question, we propose learning with Stochastic Turn Depth (**StochasT**) to harmonize LVLMs' capabilities for singleT and multiT testing. Inspired by the concept of neuron or layer dropout [24, 49], StochasT naturally stochastically varies the number of historical conversational turns for each training sample. Through a controlled expansion of the underlying conversation tree, this strategy substantially enriches the training context distribution and improves contextual diversity. Crucially, this is achieved without introducing additional data or extra computational overhead, effectively training the model on an "ensemble" of multiT data configurations. Furthermore, StochasT preserves the total number of training samples while diversifying their internal structure and depth, allowing it to be seamlessly integrated with other training objectives and novel structures [5, 41, 57, 70].

Our approach offers three major advantages: 1) **Enhanced Visual Alignment**: By effectively modulating the context window length, we improve the attention weights allocated from loss tokens to vision tokens, leading to stronger visual grounding. 2) **Diverse Turn Distribution**: The method substantially diversifies both the distribution of turn counts and the historical content models see during training, enabling the model to handle various real-world scenarios with fluctuating context lengths. 3) **Efficient Data Augmentation**: The dynamic generation of turn variants serves as a powerful, compute-neutral data augmentation strategy that enhances model generalization while preserving training efficiency. We validate our method across five distinct domain tasks [20–22, 38, 50] in both multiT and singleT evaluation settings.

Finally, we note the performance gap between singleT and multiT evaluation exposes the brittle robustness of LVLMs, posing a risk when relying on either mechanism for LVLM assessment. Hence, we propose a novel evaluation mechanism, Balanced Latin Square Turn Permutation, that broadens the scope of evaluation beyond isolated singleT or multiT setting, focusing instead on the robustness of model outputs under various contextual perturbations.

We summarize our contribution as follows:

- To the best of our knowledge, we are the first to highlight and systematically analyze the discrepancy between the common practice of multiT VIT and the prevalent singleT evaluation in the LVLM literature.
- We propose StochasT, a compute-neutral data augmentation and training strategy that dynamically varies historical context to enhance LVLMs' visual alignment and contextual robustness.

– We propose Balanced Latin Square Turn Permutation, an evaluation mechanism to assess LVLMs' robustness under diverse contextual perturbations, moving beyond traditional singleT or multiT testing.

2 Related Work

2.1 Visual Instruction Tuning

Contemporary large vision-language models (LVLMs) [2,3,13,19,33,34,51,52,58, 59,71,72] typically adopt the architecture that comprise a vision encoder [14,45, 67], a cross-modal projector (e.g., an MLP or Q-Former [30]), and a pretrained large language model (LLM) [10,25,44,54]. To elicit instruction-following capabilities over visual inputs, visual instruction tuning (VIT) [13,33,34,71] has emerged as the standard paradigm, analogous to language instruction tuning [42, 60, 61]. VIT generally proceeds in two stages: (1) *pretraining* the projector to align visual features with the text space, and (2) *fine-tuning* the model via supervised fine-tuning (SFT), often followed by reinforcement learning (RL) [42, 46, 47] for enhanced alignment. Building upon this foundational pipeline, recent advancements primarily focus on refining training data and developing novel objectives. On the data front, studies demonstrate that rigorous curation is often more critical to VIT efficacy than sheer volume [62, 69]. Alongside these curation efforts, DRESS [9] leverages natural language human feedback to enhance both alignment and multiT interaction capabilities. Beyond data-driven enhancements, several approaches modify training objectives to mitigate overfitting and improve generalization. L2T [70] introduces an auxiliary loss on instruction tokens to penalize shortcut learning and encourage the model to focus on visual tokens. Vittle [41] formulates a novel learning objective grounded in the information bottleneck principle to increase robustness against distribution shifts. Furthermore, Ross [57] augments standard text supervision with a denoising objective that reconstructs latent visual representations, an approach recently generalized to 3D representation learning by Ross3D [56]. Unlike prior works that neglect the balance between singleT and multiT capabilities during VIT, our approach naturally maintains an equilibrium without requiring auxiliary objectives or additional budget for data curation and training.

2.2 Visual Instruction Tuning Datasets & LVLM Evaluation

VIT datasets generally consist of large-scale, general-purpose corpora (e.g., LLaVA-Instruct-665K [33], MiniGPT-4 [71], SVIT [68]) and specialized downstream datasets. These specialized sets address targeted tasks such as document understanding [40], hierarchical recognition [50], biomedical VQA [12], and various other domain-specific applications [20, 22, 29, 31, 32, 43, 65]. LVLMs instruction-tuned on these data are predominantly evaluated in singleT settings using standard benchmarks like MMMU [66], MathVista [39], MMBench [37], MMStar [8], and MME [18]. Concurrently, recent research has increasingly focused on eliciting the multiT conversational capabilities of LVLMs [6, 17, 27, 38, 63], prompting

the development of benchmarks specifically tailored for multiT interactions [16, 35, 36, 38]. Despite these advancements, a critical gap remains: the absence of training recipes and balanced evaluation metrics designed to jointly foster and assess both singleT and multiT paradigms.

3 Method

3.1 SingleT vs. MultiT Visual Instruction Tuning (VIT)

A Large Vision-Language Model (LVLM), parameterized by θ, is designed to process a visual input X_v and a text instruction X_q, generating a corresponding textual response X_a of length L. This generation process is modeled autoregressively as:

$$P_\theta(X_a|X_v, X_q) = \prod_{i=1}^{L} P_\theta(x_{a,i}|X_v, X_q, x_{a,<i}), \tag{1}$$

where $x_{a,i}$ is the i-th token of the response X_a, and $x_{a,<i}$ denotes the sequence of preceding tokens.

In the context of multiT VIT, the model processes a conversational dialogue $\mathcal{C} = \{(X_q^{(n)}, X_a^{(n)})\}_{n=1}^{N}$ based on the image X_v, where N denotes the total number of dialogue turns. At the n-th turn, $X_q^{(n)}$ represents the user instruction and $X_a^{(n)}$ represents the target response. Under the standard multiT paradigm, the sequence is packed, and the objective is to minimize the negative log-likelihood of the response tokens across all N turns:

$$\mathcal{L}_{\text{multi}} = -\sum_{n=1}^{N}\sum_{i=1}^{L_n} \log P_\theta(x_{a,i}^{(n)}|X_v, H^{(n)}, x_{a,<i}^{(n)}), \tag{2}$$

where L_n is the sequence length of the n-th response, and $H^{(n)}$ denotes the accumulated conversational context prior to the n-th response, defined as $H^{(n)} = [X_q^{(1)}, X_a^{(1)}, \ldots, X_q^{(n)}]$. Pioneering frameworks, such as LLaVA [34], synthesize these N-turn dialogues $\mathcal{C}$ using advanced large language models. The generated instructions $X_q^{(n)}$ encompass a diverse array of visual tasks, ranging from object categorization and counting to spatial reasoning and action recognition. By optimizing $\mathcal{L}_{\text{multi}}$, the LVLM learns instruction-following capabilities across diverse contexts within a single forward pass.

However, because the N turns within the dialogue $\mathcal{C}$ often focus on disparate visual features with minimal semantic dependency, an alternative approach is to unroll the multiT dialogue into N independent singleT samples, defined as $\mathcal{S} = \{(X_v, X_q^{(n)}, X_a^{(n)})\}_{n=1}^{N}$. In this singleT paradigm, the historical context is discarded such that $H^{(n)} = X_q^{(n)}$, and the model optimizes an unrolled objective over the independent pairs:

$$\mathcal{L}_{\text{single}} = -\sum_{n=1}^{N}\sum_{i=1}^{L_n} \log P_\theta(x_{a,i}^{(n)}|X_v, X_q^{(n)}, x_{a,<i}^{(n)}). \tag{3}$$

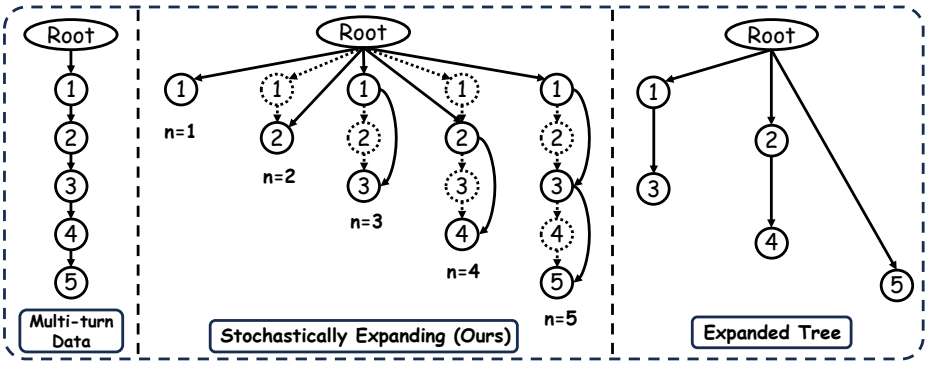

Fig. 3: An overview of our StochasT procedure. Starting from the first turn of a multiT dialogue, the method randomly omits previous context turns to generate an expanded conversation tree for the training process. The root contains the system prompt and the image grounding the turns.

While structurally simpler, duplicating X_v across N isolated samples is computationally less efficient than the packed multiT sequence. Despite this inefficiency, the singleT setting is still adopted in the training setups of some prior works [5,9] and in the organization of several datasets [20, 32].

This structural divergence raises a critical, largely overlooked research question: *How do the multiT ($\mathcal{L}_{multi}$) and singleT ($\mathcal{L}_{single}$) training paradigms differentially impact an LVLM's intrinsic capabilities and its performance on downstream tasks?* In this paper, we systematically investigate this dynamic and propose a robust visual instruction tuning strategy that harmonizes both paradigms, alongside a novel evaluation strategy utilizing two new metrics to benchmark model robustness.

3.2 Stochastic Turn Depth

Dropout [49] is a fundamental regularization technique widely adopted in deep neural networks. It randomly deactivates neurons during training, implicitly creating an ensemble of networks with varying capacities. Beyond individual neurons, this principle has been extended to entire architectural blocks. For instance, Stochastic Depth [24] randomly drops residual blocks within deep ResNets to implicitly train an ensemble of networks with varying depths, demonstrating superior performance over standard training paradigms. Building on this idea, we propose Stochastic Turn Depth (StochasT). Specifically, we perform visual instruction tuning by stochastically pruning the historical context of each turn during training. Learning with stochastic turn depth can therefore be interpreted as training an LVLM over an implicit ensemble of conversational trajectories with dynamically varying context lengths. Under this unified framework, the standard singleT and multiT training paradigms naturally arise as boundary cases.

For a training dataset $\mathcal{D}$ comprising multiT dialogues in the form of $\mathcal{C} = \{(X_q^{(n)}, X_a^{(n)})\}_{n=1}^{N}$, the most intuitive approach to implementing StochasT is to stochastically drop individual turns from each conversation. We assign a retention probability to each turn, analogous to standard dropout mechanisms. Specifically, for the n-th turn, we define a drop probability p_n sampled from a Beta distribution:

$$p_n \sim \text{Beta}(\alpha, \beta), \tag{4}$$

where α and β are hyperparameters controlling the shape of the distribution, allowing us to flexibly bias the dropout rate toward specific interaction lengths. We then sample a binary indicator mask $m_n \sim \text{Bernoulli}(1 - p_n)$ for each turn $n \in \{1, \ldots, N\}$. Therefore, after sampling, the original training dialogue $\mathcal{C}$ is reduced to a stochastically pruned conversation $\hat{\mathcal{C}}$, formally defined as:

$$\hat{\mathcal{C}} = \left\{ (X_q^{(n)}, X_a^{(n)}) \mid m_n = 1, \text{ for } n = 1, \ldots, N \right\}. \tag{5}$$

In this formulation, if $m_n = 0$, the n-th instruction-response pair is entirely removed from the sequence for the current training iteration. Consequently, the historical context H for any subsequent retained turn $k > n$ is dynamically altered.

While directly dropping simplifies data preprocessing and accelerates training by physically reducing the sequence length, it inherently decreases the number of effective tokens available for loss calculation per training sample. To maximize token utilization while maintaining context diversity, we propose learning with stochastic depth (StochasT). Instead of explicitly removing turns from the input sequence, we retain all N turns of the dialogue $\mathcal{C}$ for loss computation but stochastically mask their historical dependencies.

Algorithm 1 Stochastic Turn Depth

Input: N dialogue turns, Beta parameters α, β.
Output: Parent node assignments P.

[Optional] Randomly shuffle turns
for $n = 1$ **to** N **do**
  $k \leftarrow n - 1$
  while $k > 0$ **do**
    $p_k \sim \text{Beta}(\alpha, \beta)$,
    $m_k \sim \text{Bernoulli}(1 - p_k)$
    if m_k **then**
      break
    end if
    $k \leftarrow k - 1$
  end while
  $P[n] \leftarrow k$ {$k = 0$ connects to root X_v}
end for
return P

Due to the inherent autoregressive architecture of LVLMs, sequence interactions are governed by a causal attention mask and positional embeddings, which grant each token a unique positional index. We leverage this architecture to perform turn dropout in the backward direction, thereby preserving strict chronological causality. For each turn $n \in \{1, \ldots, N\}$, rather than assigning a deterministic, sequential history $H^{(n)}$, we stochastically connect the turn to a sampled parent node $P(n)$ to construct a diversified history $\tilde{H}^{(n)}$. Consequently, this procedure can be formulated as a directed graph expansion. We incrementally build this dialogue tree from

a root node representing all tokens prior to the first turn. By convention, we designate the initial visual input X_v as this root.

Formally, to determine the parent $P(n)$ for the n-th turn, we traverse backward through its preceding nodes $k = n-1, n-2, \ldots, 1$. At each step k, we sample a drop probability $p_k \sim \text{Beta}(\alpha, \beta)$ and an indicator variable $m_{n,k} \sim \text{Bernoulli}(1-p_k)$. The backward traversal halts at the first node where $m_{n,k} = 1$, establishing it as the parent node to which turn n connects. If all preceding nodes are dropped (i.e., $m_{n,k} = 0$ for all $k < n$), the turn connects directly to the root, effectively setting $P(n) = 0$. The resulting procedural logic for constructing this stochastic tree is detailed in Algorithm 1.

By executing this algorithm, the sequentially packed multiT dialogue expands into a structured tree graph, as illustrated in Fig. 3. Our method effectively resolves the trade-off between the singleT and multiT paradigms. Compared to the flat structure of singleT training $\mathcal{S}$, our method provides greater depth and complex contextual dependencies (preserving multiple nodes along specific branches). Conversely, compared to the rigid chain of standard multiT training $\mathcal{C}$, it introduces greater breadth and contextual diversity while strictly respecting the chronological causality of the original dialogue. Note that when the drop probability $p_n \to 1$ for all turns, StochasT naturally degenerates into the singleT paradigm $\mathcal{S}$, whereas setting $p_n = 0$ strictly recovers the standard multiT paradigm $\mathcal{C}$. Ultimately, this approach reaches a robust equilibrium, optimizing the LVLM's capabilities across varying conversational lengths. Note that StochasT is best suited for the conversations where turns are not strongly interdependent.

It is instructive to contrast our StochasT tree expansion with the well-known Chinese Restaurant Process (CRP) [53]. While both are discrete stochastic processes that incrementally assign sequential entities (turns versus customers) to previous states to form complex structural dependencies, their underlying mechanisms and objectives diverge significantly. The CRP models cluster assignment through preferential attachment. Each new customer selects a table with probability proportional to its current occupancy, without considering the temporal order of previous arrivals. This naturally yields un-ordered, disjoint partitions. In contrast, StochasT constructs a causal directed tree by enforcing a strict chronological backward traversal. Rather than relying on cluster popularity, our stochastic assignment is governed by consecutive Bernoulli trials tied to temporal proximity. This critical distinction ensures that the autoregressive causality and sequence dependency inherent to LVLMs are rigorously preserved while still injecting the desired structural stochasticity.

4 Balanced Latin Square Turn Permutations

A general-purpose LVLM should yield consistently correct responses to an instruction regardless of whether it is posed in isolation or embedded within a complex, multiT dialogue. However, Fig. 2 reveals a significant performance gap between singleT and multiT evaluation, indicating that models are highly sen-

sitive to domain shifts and contextual variations. Standard evaluation metrics, which compute raw accuracy on static question-answer pairs, fail to capture this vulnerability. While some recent works [15,22] employ circular (of choices) evaluation on multi-choice questions, directly extending it from choices to turns brings in limited contextual variation; with the original inter-turn adjacency intact.

To rigorously assess model robustness under varying contextual dependencies, we propose a novel evaluation paradigm based on the Balanced Latin Square (BLS) [4]. BLS constructs an $N \times N$ matrix where each turn appears exactly once in every dialogue position, and every turn immediately precedes every other turn exactly once. Hence, evaluating LVLMs following these permutations can systematically control for both absolute positional bias and first-order carry-over effects. A standard BLS requires N to be an even number; if N is odd, fully balancing the permutations requires a $2N \times N$ matrix, which doubles the inference budget. To maintain computational efficiency, if the number of evaluation turns N is odd, we pad the dialogue sequence with a universal placeholder instruction: $X_q^{(\text{dummy})} =$ `Please briefly describe this image.` This placeholder does not contribute to final accuray. This ensures the effective sequence length $\tilde{N}$ is always even (where $\tilde{N} = N$ if N is even, and $\tilde{N} = N + 1$ if N is odd), allowing us to construct a perfectly balanced $\tilde{N} \times \tilde{N}$ permutation matrix corresponding to exactly $\tilde{N}$ inferences.

Let $c_{n,j} \in \{0, 1\}$ denote the binary correctness of the model's response to the n-th turn when evaluated under the j-th permutation sequence. We introduce two novel metrics to quantify a model's intrinsic capabilities.

Context-Robust Accuracy (CRA). This metric computes the average accuracy of a specific turn across all $\tilde{N}$ contextual permutations,

$$\text{CRA}^{(n)} = \frac{1}{\tilde{N}} \sum_{j=1}^{\tilde{N}} c_{n,j} \quad \text{and} \quad \text{CRA} = \frac{1}{\tilde{N}} \sum_{n=1}^{\tilde{N}} \text{CRA}^{(n)}, \tag{6}$$

which measures a model's robustness under distinct contextual histories.

Strict Context-Robust Accuracy (CRA+). We then define a more stringent metric,

$$\text{CRA+}^{(n)} = \prod_{j=1}^{\tilde{N}} c_{n,j} \quad \text{CRA+} \; = \; \frac{1}{\tilde{N}} \sum_{n=1}^{\tilde{N}} \text{CRA+}^{(n)}, \tag{7}$$

where $\text{CRA+}^{(n)}$ is 1 only when the model makes no mistake under all the turn histories / permutations. It assess the model's genuine, context-invariant knowledge regarding the visual instruction.

Limitation and Discussion. The proposed evaluation mechanism assumes the multiple instruction-answer turns about an image are independent of each other, to some extent. It holds for a majority of existing LVLM evaluation benchmarks, to the best of our knowledge. Still, it is important to note that one should not apply it to conversations with strong dependency on historic contexts.

5 Experiments

We validate our proposed StochasT framework using Qwen2.5-VL-3B [3] and LLaVA-1.5-7B [33]. Following the experimental setup detailed in Sec. 5.1, we comprehensively compare our approach against standard multiT and singleT training baselines across a diverse suite of downstream tasks and leverage we our novel Balanced Latin Square metrics to assess models' robustness (Sec. 5.2). In Sec. 5.3, we conduct ablation studies on the dropout distribution hyperparameters, and systematically investigate the impact of StochasT on visual attention.

5.1 Experimental Setup

Datasets. We select diverse VIT datasets spanning multiple domains, specifically targeting downstream tasks that inherently feature multiT instructions per image. **iNat-Plant** [50] is a hierarchically constructed instruction tuning dataset designed for fine-grained visual understanding, grounded in the plant taxonomy of iNaturalist-2021 [55]. **PathVQA** [21] is a pioneering dataset focused on clinical pathology. As it lacks an explicit visual instruction tuning subset, we programmatically reformat its original training split into a standard instruction-response format. **CoralVQA** [20] is curated to develop and assess the specialized capability of LVLMs in coral reef analysis. **TaiwanVQA** [22] targets the adaptation of LVLMs to culturally specific visual content, evaluating both localized recognition and nuanced reasoning. Additionally, we validate our approach on **MMDU** [38], an explicit multiT, multi-image instruction tuning dataset designed to enhance an LVLM's extended conversational capabilities and complex multi-image reasoning.

Benchmarks and Evaluation. We evaluate the trained models on their respective official benchmarks or test splits. For the multiple-choice tasks in iNat-Plant and TaiwanVQA, we directly extract the model's predicted options. For the open-ended generative tasks in MMDU, CoralVQA, and PathVQA, we employ Gemini 3 Flash to evaluate the generated responses.

Implementation Details. We apply LoRA [23] fine-tuning to the official visual instruction-tuned checkpoints of Qwen2.5-VL and LLaVA-1.5. Our training pipeline leverages the official LLaVA-1.5 codebase and the designated fine-tuning repository [26] for Qwen2.5-VL. Comprehensive hyperparameters, hardware configurations, and training specifics are detailed in the supplement.

5.2 Main Results

We compare our method against the standard multiT visual instruction tuning baseline as well as a specialized singleT training setting. The evaluations are conducted under both singleT and multiT settings, as detailed in Tab. 1. We report the average accuracy across all four datasets, while listing MMDU separately as it does not support singleT training and evaluation. Detailed results for each individual dataset and Qwen3-VL-32B [2] are provided in the supplement.

Table 1: Performance comparison of different training strategies on LLaVA-1.5-7B and Qwen2.5-VL-3B under singleT and multiT evaluation settings.

Model	Eval.	Training	Avg. Accuracy over [20–22, 50]	MMDU [38]
LLaVA-1.5-7B	SingleT	Original	43.74	–
		MultiT	61.51	–
		SingleT	**68.46**	–
		StochasT	67.16	–
	MultiT	Original	46.84	10.81
		MultiT	68.52	12.68
		SingleT	68.60	–
		StochasT	**70.63**	**13.10**
Qwen-2.5-VL-3B	SingleT	Original	55.48	–
		MultiT	71.20	–
		SingleT	76.71	–
		StochasT	**77.28**	–
	MultiT	Original	58.16	47.00
		MultiT	77.39	57.90
		SingleT	77.75	–
		StochasT	**80.16**	**59.80**

As shown in Tab. 1, our method consistently outperforms MultiT across all evaluation settings for both models. Specifically, StochasT yields an average performance gain of 9.19% relative to singleT evaluation and 3.08% over multiT evaluations for LLaVA-1.5-7B, alongside gains of 8.54% and 3.58%, respectively, for Qwen2.5-VL-3B. As expected, SingleT also outperforms MultiT across all singleT evaluations due to the in-distribution nature of this testing. Additionally, SingleT achieves comparable performance to MultiT on multiT evaluations. We conjecture that this is caused by the effectively shorter sequence lengths in context, allowing loss tokens to attend more efficiently to the visual inputs and instructions. We discuss this dynamic in detail in the subsequent section.

Compared to SingleT, our method achieves state-of-the-art (SOTA) performance on all multiT evaluations across both models, while remaining highly competitive on singleT evaluations. Importantly, as illustrated in Fig. 4, SingleT requires approximately double the number of tokens to achieve similar singleT performance. This inefficiency arises from data unrolling, where the same image is duplicated multiple times within a single training epoch, leading to a massive influx of redundant tokens. This contrast highlights the perfect equilibrium StochasT achieves between training efficiency and downstream performance.

SingleT vs. MultiT Evaluation. The MultiT baseline exhibits a noticeable vulnerability when evaluated on isolated singleT instructions, struggling to gen-

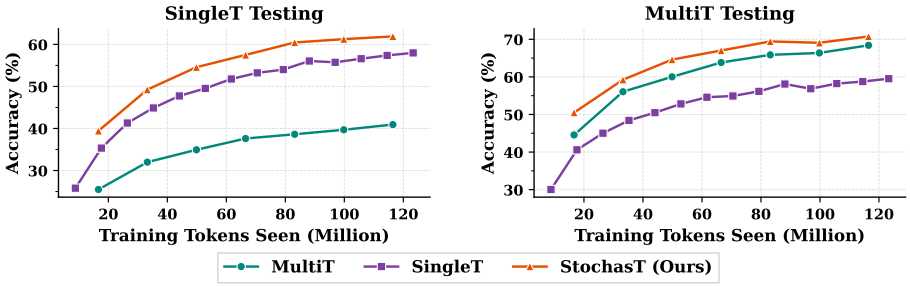

Fig. 4: Accuracy vs. #training tokens for different training strategies. We evaluate MultiT, SingleT, and our proposed StochasT under singleT and multiT testing. As the number of training tokens increases, StochasT consistently outperforms both baselines and shows faster convergence, especially in the low-token regime.

eralize outside of a continuous conversational context. multiT evaluation consistently outperforms singleT evaluation, exposing an average gap of 7.01% on LLaVA-1.5-7B and 6.19% on Qwen2.5-VL-3B. Notably, the original gap before downstream instruction tuning was less than half of this magnitude (3.10% and 2.68%, respectively). This indicates that standard multiT training inherently exacerbates this discrepancy, rendering the models increasingly brittle to contextual perturbations. With StochasT, this performance gap narrows significantly to 3.47% and 3.33%, closely resembling the baseline robustness of the pre-trained models. This demonstrates the effectiveness of our approach in maintaining original structural robustness while heavily regularizing the learned distribution. For SingleT, this gap is reduced to less than 1%, making it an ideal debiasing choice exclusively in scenarios where computational overhead is not a concern.

Table 2: The context-robust accuracy (CRA) and strict context-robust accuracy (CRA+) averaged over four datasets.

Method	CRA	CRA+
LLaVA-1.5-7B	45.73	27.66
+MultiT	61.82	41.19
+SingleT	67.33	**53.15**
+**StochasT (Ours)**	**67.69**	50.89
Qwen2.5-VL-3B	57.26	38.46
+MultiT	73.63	54.30
+SingleT	77.26	**64.57**
+**StochasT (Ours)**	**77.53**	61.76

MultiT Multi-Image Training. Surprisingly, StochasT even outperforms MultiT on MMDU, a complex long-context dialogue benchmark with related sequential turns. We conjecture that this advantage stems from the relatively weak interdependency among MMDU turns, as observed in [63]. Under such mild dependencies, StochasT remains robust by preserving the original turn order while encouraging stronger visual grounding rather than relying on brittle textual shortcuts.

Balanced Latin Square Evaluation. Table 2 presents the results of our Balanced Latin Square evaluation. On both models, StochasT achieves the highest Context-Robust Accuracy (CRA), indicating comprehensively superior perfor-

mance across diverse historical contexts. This demonstrates its immense utility as a foundational training strategy for real-world deployment, and we advocate for CRA as a primary metric for generalized LVLM evaluation. For Strict Context-Robust Accuracy (CRA+), which requires consistency under all contextual perturbations, SingleT marginally outperforms StochasT, since training with repeatedly duplicated images forces the model toward highly deterministic outputs through brute-force reinforcement. Nevertheless, this establishes CRA+ as an excellent proxy for an LVLM's predictive uncertainty.

5.3 Ablation and Analysis

Ablation on Dropout Distribution. We adopt the Beta distribution to sample dropout rates because it flexibly models bounded probabilities with only two parameters. By varying (α, β), it can smoothly interpolate between SingleT-like shallow histories and MultiT-like deep histories, making it a simple and practical default among possible bounded distributions. We compare different (α, β) settings that emphasize varying amounts of contextual history: higher dropout rates reduce history and encourage visual grounding, while lower dropout rates preserve deeper conversations for long-context reasoning. By default, all main experiments use a symmetric unimodal distribution with $\alpha = 2$ and $\beta = 2$.

We evaluate four distinct configurations on CoralVQA and iNat-Plant: a U-shaped distribution $(0.5, 0.5)$, a lower-biased distribution $(1, 5)$, a higher-biased distribution $(5, 1)$, and our default symmetric setting $(2, 2)$. As shown in Tab. 3 (Left), performance remains relatively robust across the parameter space,with the symmetric $(2, 2)$ setting achieving the best overall results. Notably, $(5, 1)$, which favors shorter SingleT-like histories, consistently outperforms $(1, 5)$, which is closer to MultiT. Overly suppressing dropout makes the framework regress toward deterministic MultiT training, reintroducing contextual overfitting and reducing the gains under the stringent CRA+ metric.

Table 3: Left: Effect of Beta parameters (α, β) on model performance. **Right**: Comparison with random-fixed cutoff (RC) baseline, averaged over two datasets.

Metric	Beta Parameters (α, β)				Qwen2.5VL-3B		
	(0.5,0.5)	(2,2)	(1,5)	(5,1)	MultiT	RC	**Ours**
CRA	80.92	**82.26**	81.04	81.98	73.65	70.61	**79.24**
CRA+	69.05	**73.30**	69.81	72.69	44.05	44.95	**56.41**

Ablation on Turn Depth. To examine whether the gains simply arise from a smaller expected turn depth, we compare StochasT with a random-fixed cutoff (RC) baseline on iNat-Plant and PathVQA, matching their expected turn depths and reporting the averaged results in Tab. 3 (Right). Directly shortening the turn depth underperforms even MultiT, suggesting that StochasT's improvements do not merely result from a shorter effective context or generic context perturbation.

Impact of StochasT on Visual Attention. In LVLMs, visual tokens typically occupy a vast portion of the attention matrix while receiving a disproportionately small fraction of the total attention weights [1, 7, 70], whereas text tokens

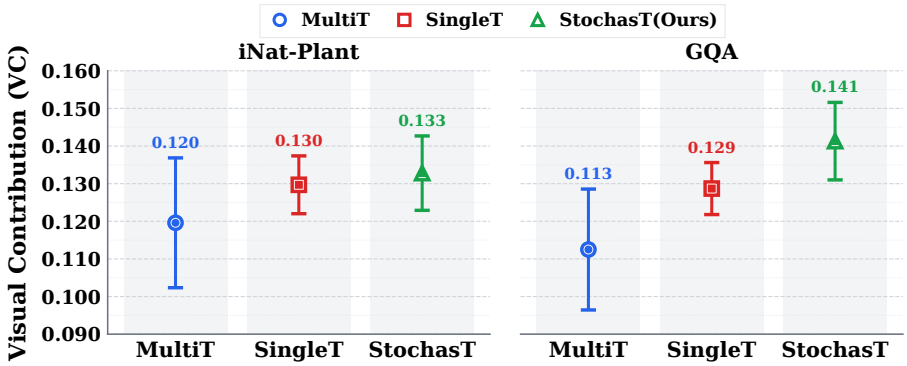

Fig. 5: Visual contribution (VC) on GQA and iNat-Plant. Mean VC and variance for MultiT, SingleT, and StochasT on iNat and GQA. Our method consistently increases visual contribution, suggesting improved visual grounding.

dominate the attention distribution. This imbalance is a widely observed bottleneck. Moreover, as the sequence length increases, this effect becomes increasingly severe, formalized by Yang *et al.* [64] as visual attention degradation.

To investigate whether StochasT effectively forces the model to better utilize visual information, we evaluate our approach using the Visual Contribution (VC) metric proposed by Zhou *et al.* [70]. VC quantifies visual reliance by measuring the difference in output negative log-likelihood between a standard forward pass and a perturbed forward pass where the image is replaced by random noise. We compute VC over 1,000 samples each from the iNat-Plant and GQA validation splits using LLaVA-1.5-7B trained on iNat-Plant.

As illustrated in Fig. 5, StochasT achieves the highest mean VC on both datasets for both task-specific and general instructions. This indicates that StochasT successfully conditions the model to anchor its generation more heavily on visual features. Additionally, MultiT exhibits the lowest mean and the highest standard deviation, rendering it highly vulnerable to input perturbations and elevating its risk of hallucination. SingleT maintains a moderate VC with a small standard deviation, indicating strong robustness to input perturbations, which firmly aligns with its high CRA+ score identified earlier.

6 Conclusion

In this work, we highlight and systematically analyze a critical, previously overlooked vulnerability in Large Vision-Language Models (LVLMs): the severe performance discrepancy arising from the structural mismatch between multiT training and singleT evaluation. To bridge this gap, we introduce StochasT, an elegant and efficient training paradigm that implicitly optimizes the model over an ensemble of dynamically varying context lengths via the stochastic dropping of conversational history. Our extensive evaluations demonstrate that StochasT

successfully harmonizes the strengths of both single- and multiT training. By penalizing reliance on spurious textual shortcuts, our method compels the LVLM to anchor its reasoning directly on the underlying visual features, thereby significantly enhancing multimodal utilization. Furthermore, we propose the Balanced Latin Square Turn Permutation evaluation framework, accompanied by two metrics: Context-Robust Accuracy (CRA) and Strict Context-Robust Accuracy (CRA+). Ultimately, these contributions not only expose the inherent brittleness of existing LVLMs but also establish a comprehensive, rigorous framework for developing inherently robust and context-resilient multimodal architectures.

Acknowledgements

This work was supported in part by NSF 2540851 and a Gemini Academic Program Award.

References

1. An, W., Tian, F., Leng, S., Nie, J., Lin, H., Wang, Q., Chen, P., Zhang, X., Lu, S.: Mitigating object hallucinations in large vision-language models with assembly of global and local attention. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 29915–29926 (2025)
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-vl technical report (2025), https://arxiv.org/abs/2511.21631
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report (2025), https://arxiv.org/abs/2502.13923
4. Bradley, J.V.: Complete counterbalancing of immediate sequential effects in a latin square design. Journal of the American Statistical Association **53**(282), 525–528 (1958), http://www.jstor.org/stable/2281872
5. Cha, J., Kang, W., Mun, J., Roh, B.: Honeybee: Locality-enhanced projector for multimodal llm. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13817–13827 (2024)
6. Chen, B., Hong, D., Ji, J., Zheng, J., Dong, B., Zhou, J., Wang, K., Dai, J., Wang, X., Chen, W., Zheng, Q., Li, W., Han, S., Guo, Y., Yang, Y.: InterMT: Multi-turn interleaved preference alignment with human feedback. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2025), https://openreview.net/forum?id=4SUtAp2cm0
7. Chen, L., Zhao, H., Liu, T., Bai, S., Lin, J., Zhou, C., Chang, B.: An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In: European Conference on Computer Vision. pp. 19–35. Springer (2024)

8. Chen, L., Li, J., Dong, X., Zhang, P., Zang, Y., Chen, Z., Duan, H., Wang, J., Qiao, Y., Lin, D., et al.: Are we on the right way for evaluating large vision-language models? Advances in Neural Information Processing Systems **37**, 27056–27087 (2024)
9. Chen, Y., Sikka, K., Cogswell, M., Ji, H., Divakaran, A.: Dress: Instructing large vision-language models to align and interact with humans via natural language feedback. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14239–14250 (2024)
10. Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., Stoica, I., Xing, E.P.: Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality (March 2023), https://lmsys.org/blog/2023-03-30-vicuna/
11. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
12. Cui, H., Mao, L., Liang, X., Zhang, J., Ren, H., Li, Q., Li, X., Yang, C.: Biomedical visual instruction tuning with clinician preference alignment. Advances in neural information processing systems **37**, 96449–96467 (2024)
13. Dai, W., Li, J., Li, D., Tiong, A., Zhao, J., Wang, W., Li, B., Fung, P.N., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning. Advances in neural information processing systems **36**, 49250–49267 (2023)
14. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
15. Duan, H., Yang, J., Qiao, Y., Fang, X., Chen, L., Liu, Y., Dong, X., Zang, Y., Zhang, P., Wang, J., et al.: Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In: Proceedings of the 32nd ACM international conference on multimedia. pp. 11198–11201 (2024)
16. Epstein, E., Yao, K., Li, J., Bai, S., Palangi, H.: Mmmt-if: A challenging multi-modal multi-turn instruction following foundation model benchmark (2024), https://arxiv.org/abs/2409.18216
17. Feng, J., Sun, Q., Xu, C., Zhao, P., Yang, Y., Tao, C., Zhao, D., Lin, Q.: Mmdialog: A large-scale multi-turn dialogue dataset towards multi-modal open-domain conversation. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 7348–7363 (2023)
18. Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., Wu, Y., Ji, R., Shan, C., He, R.: MME: A comprehensive evaluation benchmark for multimodal large language models. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2025), https://openreview.net/forum?id=DgH9YCsqWm
19. Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)
20. hongyong han, Wang, W., Zhang, G., Li, M., Wang, Y.: CoralVQA: A large-scale visual question answering dataset for coral reef image understanding. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2025), https://openreview.net/forum?id=iRsZHAMNHL
21. He, X., Zhang, Y., Mou, L., Xing, E., Xie, P.: Pathvqa: 30000+ questions for medical visual question answering. arXiv preprint arXiv:2003.10286 (2020)

22. Hsieh, H.Y., Liu, S.W., Meng, C.C., Chen, C.H., Lin, S.Y., Lin, H.J., Huang, H.H., Wu, I.C.: TaiwanVQA: Benchmarking and enhancing cultural understanding in vision-language models. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2025), https://openreview.net/forum?id=atofIc3x1q
23. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. Iclr **1**(2), 3 (2022)
24. Huang, G., Sun, Y., Liu, Z., Sedra, D., Weinberger, K.Q.: Deep networks with stochastic depth. In: European conference on computer vision. pp. 646–661. Springer (2016)
25. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L.R., Lachaux, M.A., Stock, P., Scao, T.L., Lavril, T., Wang, T., Lacroix, T., Sayed, W.E.: Mistral 7b (2023), https://arxiv.org/abs/2310.06825
26. Lee, Y.: Qwen2-vl-finetune (2024), https://github.com/2U1/Qwen2-VL-Finetune
27. Lei, Y., Yang, Z., Liu, Z., Leng, H., Liu, S., Gao, T., Liu, Q., Wang, Y.: Contextqformer: A new context modeling method for multi-turn multi-modal conversations. arXiv preprint arXiv:2505.23121 (2025)
28. Li, B., Ge, Y., Chen, Y., Ge, Y., Zhang, R., Shan, Y.: Seed-bench-2-plus: Benchmarking multimodal large language models with text-rich visual comprehension. arXiv preprint arXiv:2404.16790 (2024)
29. Li, J., Ma, W., Li, X., Lou, Y., Zhou, G., Zhou, X.: Cad-llama: Leveraging large language models for computer-aided design parametric 3d model generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18563–18573 (June 2025)
30. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)
31. Lin, W., Mirza, M.J., Doveh, S., Feris, R., Giryes, R., Hochreiter, S., Karlinsky, L.: Comparison visual instruction tuning. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2973–2983 (2025)
32. Liu, F., Lin, K., Li, L., Wang, J., Yacoob, Y., Wang, L.: Mitigating hallucination in large multi-modal models via robust instruction tuning. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=J44HfH4JCg
33. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 26296–26306 (June 2024)
34. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
35. Liu, J., Zheng, S., Karlsson, B.F., Lu, Z.: Taking notes brings focus? towards multi-turn multimodal dialogue learning. In: Christodoulopoulos, C., Chakraborty, T., Rose, C., Peng, V. (eds.) Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 33303–33324. Association for Computational Linguistics, Suzhou, China (Nov 2025). https://doi.org/10.18653/v1/2025.emnlp-main.1690, https://aclanthology.org/2025.emnlp-main.1690/
36. Liu, S., Ying, K., Zhang, H., Yang, Y., Lin, Y., Zhang, T., Li, C., Qiao, Y., Luo, P., Shao, W., et al.: Convbench: A multi-turn conversation evaluation benchmark with hierarchical ablation capability for large vision-language models. Advances in Neural Information Processing Systems **37**, 100734–100782 (2024)

37. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al.: Mmbench: Is your multi-modal model an all-around player? In: European conference on computer vision. pp. 216–233. Springer (2024)
38. Liu, Z., Chu, T., Zang, Y., Wei, X., Dong, X., Zhang, P., Liang, Z., Xiong, Y., Qiao, Y., Lin, D., et al.: Mmdu: A multi-turn multi-image dialog understanding benchmark and instruction-tuning dataset for lvlms. Advances in Neural Information Processing Systems **37**, 8698–8733 (2024)
39. Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., Cheng, H., Chang, K.W., Galley, M., Gao, J.: Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. arXiv preprint arXiv:2310.02255 (2023)
40. Luo, C., Shen, Y., Zhu, Z., Zheng, Q., Yu, Z., Yao, C.: Layoutllm: Layout instruction tuning with large language models for document understanding. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15630–15640 (2024)
41. Oh, C., Li, J., Im, S., Li, S.: Visual instruction bottleneck tuning. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025), https://openreview.net/forum?id=yzHiEmLSk8
42. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., Lowe, R.: Training language models to follow instructions with human feedback (2022), https://arxiv.org/abs/2203.02155
43. Pi, R., Zhang, J., Han, T., Zhang, J., Pan, R., Zhang, T.: Personalized visual instruction tuning. arXiv preprint arXiv:2410.07113 (2024)
44. Qwen, :, Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M., Zhang, P., Zhu, Q., Men, R., Lin, R., Li, T., Tang, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Wan, Y., Liu, Y., Cui, Z., Zhang, Z., Qiu, Z.: Qwen2.5 technical report (2025), https://arxiv.org/abs/2412.15115
45. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision (2021), https://arxiv.org/abs/2103.00020
46. Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S., Finn, C.: Direct preference optimization: Your language model is secretly a reward model. Advances in neural information processing systems **36**, 53728–53741 (2023)
47. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)
48. Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., et al.: Openai gpt-5 system card. arXiv preprint arXiv:2601.03267 (2025)
49. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research **15**(1), 1929–1958 (2014)
50. Tan, Y., Qing, Y., Gong, B.: Vision llms are bad at hierarchical visual understanding, and llms are the bottleneck (2025), https://arxiv.org/abs/2505.24840
51. Team, G., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., Rouillard, L., Mesnard, T., Cideron,

G., bastien Grill, J., Ramos, S., Yvinec, E., Casbon, M., Pot, E., Penchev, I., Liu, G., Visin, F., Kenealy, K., Beyer, L., Zhai, X., Tsitsulin, A., Busa-Fekete, R., Feng, A., Sachdeva, N., Coleman, B., Gao, Y., Mustafa, B., Barr, I., Parisotto, E., Tian, D., Eyal, M., Cherry, C., Peter, J.T., Sinopalnikov, D., Bhupatiraju, S., Agarwal, R., Kazemi, M., Malkin, D., Kumar, R., Vilar, D., Brusilovsky, I., Luo, J., Steiner, A., Friesen, A., Sharma, A., Sharma, A., Gilady, A.M., Goedeckemeyer, A., Saade, A., Feng, A., Kolesnikov, A., Bendebury, A., Abdagic, A., Vadi, A., György, A., Pinto, A.S., Das, A., Bapna, A., Miech, A., Yang, A., Paterson, A., Shenoy, A., Chakrabarti, A., Piot, B., Wu, B., Shahriari, B., Petrini, B., Chen, C., Lan, C.L., Choquette-Choo, C.A., Carey, C., Brick, C., Deutsch, D., Eisenbud, D., Cattle, D., Cheng, D., Paparas, D., Sreepathihalli, D.S., Reid, D., Tran, D., Zelle, D., Noland, E., Huizenga, E., Kharitonov, E., Liu, F., Amirkhanyan, G., Cameron, G., Hashemi, H., Klimczak-Plucińska, H., Singh, H., Mehta, H., Lehri, H.T., Hazimeh, H., Ballantyne, I., Szpektor, I., Nardini, I., Pouget-Abadie, J., Chan, J., Stanton, J., Wieting, J., Lai, J., Orbay, J., Fernandez, J., Newlan, J., yeong Ji, J., Singh, J., Black, K., Yu, K., Hui, K., Vodrahalli, K., Greff, K., Qiu, L., Valentine, M., Coelho, M., Ritter, M., Hoffman, M., Watson, M., Chaturvedi, M., Moynihan, M., Ma, M., Babar, N., Noy, N., Byrd, N., Roy, N., Momchev, N., Chauhan, N., Sachdeva, N., Bunyan, O., Botarda, P., Caron, P., Rubenstein, P.K., Culliton, P., Schmid, P., Sessa, P.G., Xu, P., Stanczyk, P., Tafti, P., Shivanna, R., Wu, R., Pan, R., Rokni, R., Willoughby, R., Vallu, R., Mullins, R., Jerome, S., Smoot, S., Girgin, S., Iqbal, S., Reddy, S., Sheth, S., Põder, S., Bhatnagar, S., Panyam, S.R., Eiger, S., Zhang, S., Liu, T., Yacovone, T., Liechty, T., Kalra, U., Evci, U., Misra, V., Roseberry, V., Feinberg, V., Kolesnikov, V., Han, W., Kwon, W., Chen, X., Chow, Y., Zhu, Y., Wei, Z., Egyed, Z., Cotruta, V., Giang, M., Kirk, P., Rao, A., Black, K., Babar, N., Lo, J., Moreira, E., Martins, L.G., Sanseviero, O., Gonzalez, L., Gleicher, Z., Warkentin, T., Mirrokni, V., Senter, E., Collins, E., Barral, J., Ghahramani, Z., Hadsell, R., Matias, Y., Sculley, D., Petrov, S., Fiedel, N., Shazeer, N., Vinyals, O., Dean, J., Hassabis, D., Kavukcuoglu, K., Farabet, C., Buchatskaya, E., Alayrac, J.B., Anil, R., Dmitry, Lepikhin, Borgeaud, S., Bachem, O., Joulin, A., Andreev, A., Hardin, C., Dadashi, R., Hussenot, L.: Gemma 3 technical report (2025), https://arxiv.org/abs/2503.19786

52. Team, K., Du, A., Yin, B., Xing, B., Qu, B., Wang, B., Chen, C., Zhang, C., Du, C., Wei, C., et al.: Kimi-vl technical report. arXiv preprint arXiv:2504.07491 (2025)
53. Teh, Y., Jordan, M., Beal, M., Blei, D.: Sharing clusters among related groups: Hierarchical dirichlet processes. Advances in neural information processing systems **17** (2004)
54. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: Llama: Open and efficient foundation language models (2023), https://arxiv.org/abs/2302.13971
55. Van Horn, G., Cole, E., Beery, S., Wilber, K., Belongie, S., Mac Aodha, O.: Benchmarking representation learning for natural world image collections. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12884–12893 (2021)
56. Wang, H., Zhao, Y., Wang, T., Fan, H., Zhang, X., Zhang, Z.: Ross3d: Reconstructive visual instruction tuning with 3d-awareness. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9275–9286 (2025)

57. Wang, H., Zheng, A., Zhao, Y., Wang, T., Ge, Z., Zhang, X., Zhang, Z.: Reconstructive visual instruction tuning. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=8q9NOMzRDg
58. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
59. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., Wang, Z., Chen, Z., Zhang, H., Yang, G., Wang, H., Wei, Q., Yin, J., Li, W., Cui, E., Chen, G., Ding, Z., Tian, C., Wu, Z., Xie, J., Li, Z., Yang, B., Duan, Y., Wang, X., Hou, Z., Hao, H., Zhang, T., Li, S., Zhao, X., Duan, H., Deng, N., Fu, B., He, Y., Wang, Y., He, C., Shi, B., He, J., Xiong, Y., Lv, H., Wu, L., Shao, W., Zhang, K., Deng, H., Qi, B., Ge, J., Guo, Q., Zhang, W., Zhang, S., Cao, M., Lin, J., Tang, K., Gao, J., Huang, H., Gu, Y., Lyu, C., Tang, H., Wang, R., Lv, H., Ouyang, W., Wang, L., Dou, M., Zhu, X., Lu, T., Lin, D., Dai, J., Su, W., Zhou, B., Chen, K., Qiao, Y., Wang, W., Luo, G.: Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency (2025), https://arxiv.org/abs/2508.18265
60. Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N.A., Khashabi, D., Hajishirzi, H.: Self-instruct: Aligning language models with self-generated instructions. In: Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers). pp. 13484–13508 (2023)
61. Wei, J., Bosma, M., Zhao, V.Y., Guu, K., Yu, A.W., Lester, B., Du, N., Dai, A.M., Le, Q.V.: Finetuned language models are zero-shot learners (2022), https://arxiv.org/abs/2109.01652
62. Wei, L., Jiang, Z., Huang, W., Sun, L.: Instructiongpt-4: A 200-instruction paradigm for fine-tuning minigpt-4 (2023), https://arxiv.org/abs/2308.12067
63. Yan, D., Li, Y., Chen, Q.G., Luo, W., Wang, P., Zhang, H., Shen, C.: Mmcr: Advancing visual language model in multimodal multi-turn contextual reasoning. arXiv preprint arXiv:2503.18533 (2025)
64. Yang, J., Cui, C., Zhou, Y., Chen, Y., Xia, P., Wei, Y., Yu, T., Huang, Y., Wang, L.: Ikod: Mitigating visual attention degradation in large vision-language models. arXiv preprint arXiv:2508.03469 (2025)
65. Yuan, Y., Li, W., Liu, J., Tang, D., Luo, X., Qin, C., Zhang, L., Zhu, J.: Osprey: Pixel understanding with visual instruction tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 28202–28211 (2024)
66. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., et al.: Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9556–9567 (2024)
67. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training (2023), https://arxiv.org/abs/2303.15343
68. Zhao, B., Wu, B., He, M., Huang, T.: Svit: Scaling up visual instruction tuning (2023), https://arxiv.org/abs/2307.04087
69. Zhou, C., Liu, P., Xu, P., Iyer, S., Sun, J., Mao, Y., Ma, X., Efrat, A., Yu, P., Yu, L., et al.: Lima: Less is more for alignment. Advances in Neural Information Processing Systems **36**, 55006–55021 (2023)

70. Zhou, Z., Hong, F., Luo, J., Ye, Y., Yao, J., Li, D., Han, B., Zhang, Y., Wang, Y.: Learning to instruct for visual instruction tuning. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025), https://openreview.net/forum?id=NQSWkmjODD
71. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)
72. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., Gao, Z., Cui, E., Wang, X., Cao, Y., Liu, Y., Wei, X., Zhang, H., Wang, H., Xu, W., Li, H., Wang, J., Deng, N., Li, S., He, Y., Jiang, T., Luo, J., Wang, Y., He, C., Shi, B., Zhang, X., Shao, W., He, J., Xiong, Y., Qu, W., Sun, P., Jiao, P., Lv, H., Wu, L., Zhang, K., Deng, H., Ge, J., Chen, K., Wang, L., Dou, M., Lu, L., Zhu, X., Lu, T., Lin, D., Qiao, Y., Dai, J., Wang, W.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models (2025), https://arxiv.org/abs/2504.10479