Looking Beyond Language Priors: Enhancing Visual Comprehension and Attention in Multimodal Models*

Aarti Ghatkesar
aarti@cerebras.net

Uddeshya Upadhyay
uddeshya.upadhyay@cerebras.net

Ganesh Venkatesh
ganesh.venkatesh@cerebras.net

APPLIEDML, CEREBAS

Abstract

Achieving deep alignment between vision and language remains a central challenge for Multimodal Large Language Models (MLLMs). These models often fail to fully leverage visual input, defaulting to strong language priors. Our approach first provides insights into how MLLMs internally build visual understanding of image regions and then introduces techniques to amplify this capability. We explore techniques designed both to deepen the model's understanding of visual content and to ensure that these visual insights actively guide language generation. We demonstrate the superior multimodal understanding of our resultant model through a detailed upstream analysis on predicting visually-dependent tokens as well as >10 percentage point boost on a visually challenging task and a consistent boost across multiple tasks.

1. Introduction

Recent years have witnessed remarkable progress in the field of multimodal artificial intelligence, particularly with the advent of Large Multimodal Models (MLLMs) [1, 5, 9, 21]. These models, capable of jointly processing information from different modalities like vision and language, have demonstrated impressive general-purpose capabilities across a wide range of tasks. This success has spurred significant interest and research, pushing the boundaries of what AI systems can perceive, understand, and communicate about the visual world.

Despite these striking capabilities, recent work on probing the workings of MLLMs have begun to uncover significant limitations in how current MLLMs integrate visual information [3, 11, 19]. This growing body of evidence reveals that many state-of-the-art models exhibit a strong reliance on learned language priors and statistical correlations present in their vast training datasets. Consequently, they can be surprisingly weak at leveraging fine-grained visual signals crucial for nuanced understanding. Models may generate plausible-sounding but factually incorrect responses when faced with visual details that contradict common knowledge or require careful observation, effectively 'hallucinating' or ignoring specific visual evidence in favor of more probable textual outputs. This gap highlights a critical challenge: ensuring that visual input genuinely grounds the model's reasoning and generation processes.

To address these fundamental challenges, this paper proposes a novel methodology centered on enhancing the model's internal visual representation. First, we strengthen the visual signal processing, enabling the model to extract richer and more fine-grained visual representations. Second, we introduce a training mechanism that explicitly encourages the model to allocate greater attention to these enhanced visual inputs during the response generation phase, reducing over-reliance on language priors. Finally, complementing these training modifications, we construct a targeted synthetic dataset specifically designed to leverage this advanced training fabric, providing controlled examples that force the model to learn and exploit fine-grained visual cues effectively. Our key contributions are:

Enhanced Visual Representation using VISUALLOSS. We introduce a new loss function to ensure the LLM backbone builds a rich representation of the input image, independent of whether they are mentioned in associated training text.

Weakening Language-prior via BLANKTOKENS. To fully leverage the model's enhanced visual understanding, we introduce a technique to gently reduce the LLM backbone's reliance on strong language priors. Complementing this, we develop a specialized synthetic dataset, specifically to encourage sensitivity to fine-grained visual details.

Analysis of Multimodal Alignment. We provide insights into how MLLMs attempt to understand visual information and motivates our approach. Our analysis demonstrates significant improvements in multimodal alignment, showcasing that our work pushes MLLMs towards a more visually faithful reasoning and generation.

*Accepted to the Second Workshop on Visual Concepts at CVPR ([2025]).

2. Challenges and Related Work

This section first overviews the key challenges in multimodal LLMs that we seek to address in our work and then discuss prior work on addressing these challenges.

2.1. Challenges in Multimodal LLMs

We highlight three key challenges from prior works and address them in this paper:

Weak Visual Understanding. First, several studies highlight the often weak visual understanding capabilities inherent in many contemporary MLLMs [3, 18]. This limitation is frequently attributed, at least in part, to the nature of the vision encoders employed, such as CLIP. While effective for capturing global image-text semantics, encoders like CLIP may not provide the sufficiently detailed, fine-grained representations required for nuanced visual comprehension, leading to models struggling with specific object attributes, states, or intricate scene details.

Sparse Training Signals. Second fundamental challenge limiting the depth of visual understanding in current MLLM arises from the sparse nature of the loss signals. The standard next-token prediction loss is calculated exclusively on the text sequence which provides only an indirect and often weak signal for visual learning, as only a fraction of the text tokens may have a strong, unambiguous dependence on the visual content. Furthermore, the textual descriptions frequently refer only to a subset of the objects, attributes, and relationships present in the image. As a result of this sparse supervision derived from text, the LLM backbone has limited opportunity to develop a truly rich, comprehensive internal representation of the full visual context.

Strong Language Priors. As demonstrated in prior work like SpatialEval [19], models may fail to utilize visual context correctly, particularly for tasks demanding spatial reasoning or precise grounding of textual concepts in the image. This indicates that even if relevant visual information is encoded, the mechanisms for cross-modal interaction and reasoning within the LLM component may be insufficient to properly access, interpret, and apply that information when answering questions or generating descriptions.

2.2. Related Work

Prior work attempts to address the above challenges in multiple different ways. These include strengthening visual signals by providing embeddings from multiple visual encoders to the LLM backbone [17, 18], adding auxiliary losses to help the LLM build a richer visual representation [22], and generating synthetic or augmented data [13, 16]. Our training innovations fundamentally differ from these approaches in that we do not require extra visual annotations or the overhead of additional encoders at inference time (details in Appendix 6.5).

Figure 1. Sample demonstrates how MLLMs attempt to build an internal representation that capture semantic information about the image patches even without any explicit supervision

3. Our Approach

We first analyze the internal visual understanding in existing MLLMs (Section 3.1) and rest of this section details our proposed training strategy shown in Algorithm 1.

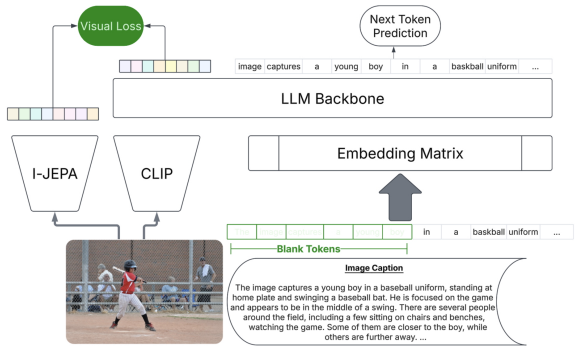
3.1. Dissecting visual tokens in MLLMs

Understanding how MLLMs internally develop visual representations is crucial for addressing their limitations and guiding future improvements. To gain insights into this process, we perform probing analyses focused specifically on the visual tokens processed by the model following initial encoding (Figure 1). In particular, we feed the visual token embeddings generated by the LLM backbone through its final layer, i.e., language modeling head (`lm_head`) and identify the top predicted token for each visual patch. We qualitatively examined a diverse set of images and present examples in Figure 1. Some key observations include:

- The model captures information about both foreground and background elements.
- The model is capable of localizing fine-grained structures or small items (e.g., `lines`, `cones`).
- The model captures *semantically distinct* items at varying scales (e.g., `handle`, `canoe`, `cow`) within an image.

Our investigations indicate that MLLMs do not treat visual tokens merely as abstract feature vectors. Instead, they exhibit an emergent capability to associate meaningful semantic labels or concepts with their corresponding image patches. This suggests that the model performs a degree of localized scene interpretation prior to extensive cross-modal fusion. Notably, MLLMs develop this localized understanding without explicit supervision for this capability.

This observation – that MLLMs develop rudimentary semantic grounding at the visual token level – provides strong motivation for our proposed approach. Given that the model possesses this nascent ability, interventions designed both to enhance this specific capability and to encourage the model to more effectively utilize this internal visual information hold significant potential. Our approach to achieving this goal is detailed in the following three sub-sections.



Innovation #1: Visual Loss. LLM backbone gets to learn dense visual concepts beyond those captured in the associated text. For example, player is wearing a red shirt, helmet, stripped pants etc.

Innovation #2: Blank Tokens. LLM backbone is required to form a coherent textual description without relying on language priors. This teaches the model to better leverage visual information.

Figure 2. Proposed modifications to MLLM training that enhances visual understanding through VISUALLOSS (Section 3.2) and encourages LLM backbone to pay greater attention to them using BLANKTOKENS (Section 3.3).

3.2. Visual Representation using VISUALLOSS

We introduce a fundamentally new strategy (referred to as VISUALLOSS in this work) to enrich the LLM's internal processing rather than just augmenting its inputs, as illustrated in Figure 2. In particular, we introduce an auxiliary vision encoder with rich visual semantic understanding (I-JEPA [2] in our work) during training phase and use it to introduce a new loss term, VISUALLOSS, on the visual tokens of the LLM backbone. By requiring the LLM backbone to predict corresponding I-JEPA representations, we provide a strong, targeted supervisory signal that directly fosters a deeper and more nuanced visual understanding within the language model itself. This process encourages the model to build a comprehensive understanding of the visual content, rather than being limited to learning only those concepts explicitly captured in the text description. Please refer to Appendix 6.1 for details.

3.3. Language Priors using BLANKTOKENS

Achieving accurate and visually grounded responses requires addressing another key challenge: over-reliance of MLLMs on strong language priors. Even with improved visual feature processing, models often default to generating text driven by learned linguistic patterns, potentially overlooking specific visual evidence. To mitigate this tendency, we introduce a complementary strategy focused on gently recalibrating the model's dependence on textual context during training. This involves strategically masking portions of the input text, thereby disrupting straightforward language-based auto-completion and compelling the model to rely more heavily on its visual comprehension capabilities to generate coherent outputs. Please refer to Appendix 6.2 for details.

Algorithm 1 Forward pass with proposed enhancements.

INPUT: I, T_{in}, T_{out}, MLLM, V_{aux}, β
OUTPUT: LOSS
#I: Image, T_{in}, T_{out}: Text input and output
#V_{aux}: Aux Vision Encoder
$\mathrm{T}_{inBlank}$ = BlankInputsPartial(T_{in}) *#Section 3.3*
V_{feat}, T_{feat} = MLLM(I, T_{in})
#V_{feat}, T_{feat} is visual and text features
V_{emb} = V_{aux}(I) *#Section 3.2*
$\mathcal{L}_{ntp} = \mathcal{CE}(\mathrm{lm\_head}(\mathrm{T}_{feat}), \mathrm{T}_{out})$ *#Next token loss*
$\mathcal{L}_{visualLoss}$ = VisualLoss(V_{feat}, V_{emb}) *#Section 3.2*
$\mathcal{L}_{tot} = \mathcal{L}_{ntp} + \beta \cdot \mathcal{L}_{visualLoss}$

3.4. Targeted Synthetic Data Generation

Figure 3. Examples of synthetic grid data for spatial awareness. Objects are randomly sampled from a large public image collection [4, 6, 7], and placed onto distinct locations.

We design synthetic training samples with the goal of further encouraging the model to focus on specific spatial relationships and visual attribute identification. We construct each image as a grid of objects using real-world natural images along with their segmentation masks from OpenImages-v7 [4, 6, 7] dataset, inspired by prior work using synthetic data to build foundational capabilities in models [16]. Figure 3 shows examples of synthetic grid visual data and Appendix 6.3 show examples of corresponding questions and additional details.

4. Results

This section analyzes the effectiveness of our proposed approach. We layer our techniques on top of a LlaVA [9] model with Llama 3.1 8B backbone (`Baseline`) and detail the training setup in Appendix 6.4. We demonstrate that our approach improves the model's ability to predict visually relevant tokens (Section 4.1) and significantly enhances its performance on challenging visual tasks (Section 4.2).

4.1. Upstream Analysis

To evaluate the impact of our techniques on the model's ability to integrate visual information during language generation, we conduct an upstream analysis on the SpatialMM dataset [15], that by construction has text with a high dependence on specific visual content, demanding strong visual

Table 1. Next token prediction (NTP) loss on SpatialMM dataset. We see an improvement as our techniques are introduced. Note that each line is additive, e.g., +BLANKTOKENS captures impact of adding VISUALLOSS and BLANKTOKENS over the baseline.

Technique	NTP
Baseline	3.59
+ VISUALLOSS	3.54
+ BLANKTOKENS	3.48
+ ind. weights	3.20
+ AIMv2 encoder	2.96
+ VISUALLOSSADV	2.94

grounding, making it suitable for this analysis.

Table 1 shows consistent decrease in cross entropy loss for the next token prediction, especially with our proposed enhancements, indicating improved prediction capability. Qualitative visualizations of token-level prediction losses (Appendix 6.6) further support this finding, showing lower values for tokens referencing visual objects, attributes, or relationships within the image.

4.2. Evaluation on SpatialEval Benchmark

We evaluate our proposed techniques on the SpatialEval [19] benchmark[1], designed to probe spatial understanding capabilities. Our results, summarized in Table 2, demonstrate a clear trend of performance improvement across all benchmark subsets with our proposal.

The addition of VISUALLOSS results in a significant improvement in model performance for `SpatialMap` and `SpatialReal` subsets of the benchmark. We attribute these improvements to VISUALLOSS's ability to strengthen the model's understanding of visual content. Adding synthetic data provides consistent improvement across all benchmark subsets, despite minimal overlap with the benchmark tasks. In particular, it helps us narrow the performance gap with the Llama 3.2 11B model [5], even though our model is smaller (8B vs 11B), utilizes a lower input resolution (e.g., 224x224 vs Llama's higher resolution of 560x560), and employs a simpler modality combination approach (input-layer feature appending vs multi-layer cross-attention). This underscores the synthetic data's effectiveness in enhancing visual grounding by leveraging our proposed training enhancements (Section 3).

4.3. Future Extensions

We explore two extensions to further improve LLM's internal visual representation. First, to mitigate the potential representational conflict arising from tasking the same LLM weights with processing both visual and textual concepts,

[1]we had to enhance benchmark's prompting and parsing capabilities which we open source as well for reproducibility

Table 2. Our proposed techniques consistently improve SpatialEvals accuracy across all benchmark subsets. `Avg.` is calculated as mean of normalized accuracy improvement.

Technique	Grid	MazeNav	Map	Real	Avg.
Baseline	34.5	15.1	44.1	50.4	1
+ VISUALLOSS	31.3	17.0	50.7	56.3	1.075
+ BLANKTOKENS	30.1	22.9	47.1	52.6	1.125
+ Synthetic	41.1	26.3	55.2	53.3	1.31
+ VISUALLOSSADV	43.5	26.7	64.1	53.3	1.385
Llama 3.2 11B [5]	50.1	28.1	48.2	53.3	1.365

we experimented with providing each modality with an independent set of weights [8, 20]. Cross-modal interaction was maintained through the standard self-attention mechanism. Our initial results in Table 1 (`+ind. weights` row) indicate that this separation significantly enhances visual representation - providing the biggest reduction in upstream loss thus far ($3.48 \Rightarrow 3.20$).

Next, we demonstrate that VISUALLOSS can leverage advancements in image encoders to help the LLM backbone build richer visual representations. We replace I-JEPA [2] with the AM-Radio [12] encoder as the auxiliary image encoder for our visual loss (shown as VISUALLOSSADV in Table 2) yields an additional gain of approximately 9.1 percentage points on the `SpatialMap` subset. This enhancement enables our model to match the performance of Llama 3.2 11B on this task, despite our model being smaller, employing a simpler architecture, and using a lower-resolution visual input. Further investigation of these extensions constitute key directions for our subsequent research.

5. Conclusion

This paper addresses the challenge of weak visual grounding in Multimodal Large Language Models (MLLMs), which often underutilize visual input due to over-reliance on language priors. We provide insights into MLLMs' nascent internal visual representation and propose novel training strategies to strengthen the LLM backbone's visual representation and its ability to leverage this understanding during response generation. The effectiveness of our approach is demonstrated through upstream analysis on the visually rich SpatialMM dataset and accuracy evaluations on the challenging SpatialEvals benchmark. We observe consistent improvement as we layer in our innovations which confirms model's superior visual reasoning capabilities. These results validate our strategy, highlighting the value of simultaneously enhancing core visual understanding and encouraging cross-modal attention as well as paves a scalable path for leveraging advances in image encoders to enhance MLLM's internal visual representation.

References

[1] Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, Soham Ghosh, Amélie Héliou, Paul Jacob, Albert Q. Jiang, Kartik Khandelwal, Timothée Lacroix, Guillaume Lample, Diego Las Casas, Thibaut Lavril, Teven Le Scao, Andy Lo, William Marshall, Louis Martin, Arthur Mensch, Pavankumar Muddireddy, Valera Nemychnikova, Marie Pellat, Patrick Von Platen, Nikhil Raghuraman, Baptiste Rozière, Alexandre Sablayrolles, Lucile Saulnier, Romain Sauvestre, Wendy Shang, Roman Soletskyi, Lawrence Stewart, Pierre Stock, Joachim Studnia, Sandeep Subramanian, Sagar Vaze, Thomas Wang, and Sophia Yang. Pixtral 12b, 2024. 1

[2] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 15619–15629, 2023. 3, 4, 1

[3] Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey. arXiv preprint arXiv:2404.18930, 2024. 1, 2

[4] Rodrigo Benenson and Vittorio Ferrari. From colouring-in to pointillism: revisiting semantic segmentation supervision. arXiv preprint arXiv:2210.14142, 2022. 3

[5] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. arXiv preprint arXiv:2407.21783, 2024. 1, 4

[6] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. International journal of computer vision, 2020. 3

[7] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. International journal of computer vision, 128(7):1956–1981, 2020. 3, 1

[8] Weixin Liang, Lili Yu, Liang Luo, Srinivasan Iyer, Ning Dong, Chunting Zhou, Gargi Ghosh, Mike Lewis, Wen-tau Yih, Luke Zettlemoyer, et al. Mixture-of-transformers: A sparse and scalable architecture for multi-modal foundation models. arXiv preprint arXiv:2411.04996, 2024. 4

[9] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. Advances in neural information processing systems, 36:34892–34916, 2023. 1, 3

[10] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International conference on machine learning, pages 8748–8763. PmLR, 2021. 1

[11] Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. Vision language models are blind. In Proceedings of the Asian Conference on Computer Vision (ACCV), pages 18–34, 2024. 1

[12] Mike Ranzinger, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. Am-radio: Agglomerative vision foundation model reduce all domains into one. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 12490–12500, 2024. 4

[13] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M. Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S. Khan. Glamm: Pixel grounding large multimodal model. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 13009–13018, 2024. 2, 3

[14] Min Shi, Fuxiao Liu, Shihao Wang, Shijia Liao, Subhashree Radhakrishnan, Yilin Zhao, De-An Huang, Hongxu Yin, Karan Sapra, Yaser Yacoob, Humphrey Shi, Bryan Catanzaro, Andrew Tao, Jan Kautz, Zhiding Yu, and Guilin Liu. Eagle: Exploring the design space for multimodal llms with mixture of encoders, 2025. 2

[15] Fatemeh Shiri, Xiao-Yu Guo, Mona Far, Xin Yu, Reza Haf, and Yuan-Fang Li. An empirical analysis on spatial reasoning capabilities of large multimodal models. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 21440–21455, 2024. 3

[16] Yihong Tang, Ao Qu, Zhaokai Wang, Dingyi Zhuang, Zhaofeng Wu, Wei Ma, Shenhao Wang, Yunhan Zheng, Zhan Zhao, and Jinhua Zhao. Sparkle: Mastering basic spatial capabilities in vision language models elicits generalization to composite spatial reasoning. arXiv preprint arXiv:2410.16162, 2024. 2, 3

[17] Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. Advances in Neural Information Processing Systems, 37:87310–87356, 2024. 2

[18] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9568–9578, 2024. 2

[19] Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, Yixuan Li, and Neel Joshi. Is a picture worth a thousand words? delving into spatial reasoning for vision language models. In Advances in Neural Information Processing Systems, pages 75392–75421. Curran Associates, Inc., 2024. 1, 2, 4

[20] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Song XiXuan, et al. Cogvlm: Visual expert for pretrained language models. Advances in Neural Information Processing Systems, 37:121475–121499, 2024. 4

[21] Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. Mm-llms: Recent advances in multimodal large language models. arXiv preprint arXiv:2401.13601, 2024. 1

[22] Hao Zhang, Hongyang Li, Feng Li, Tianhe Ren, Xueyan Zou, Shilong Liu, Shijia Huang, Jianfeng Gao, Leizhang, Chunyuan Li, et al. Llava-grounding: Grounded visual chat with large multimodal models. In European Conference on Computer Vision, pages 19–35. Springer, 2024. 2, 3

[23] Haotian Zhang, Haoxuan You, Philipp Dufter, Bowen Zhang, Chen Chen, Hong-You Chen, Tsu-Jui Fu, William Yang Wang, Shih-Fu Chang, Zhe Gan, et al. Ferret-v2: An improved baseline for referring and grounding with large language models. arXiv preprint arXiv:2404.07973, 2024. 3

Looking Beyond Language Priors: Enhancing Visual Comprehension and Attention in Multimodal ModelsAccepted to the Second Workshop on Visual Concepts at CVPR ([2025]).

Supplementary Material

6. Appendix

6.1. VisualLoss Formulation

Formally, we introduce an auxiliary vision encoder (say, $\mathcal{A}(\cdot;\theta_A)$) in addition the the standard vision encoder (defined as $\mathcal{G}(\cdot;\theta_G)$ representing CLIP [10] in our implementation connected to LLM backbone using MLP connector $\mathcal{M}$). We choose the auxiliary vision encoder to be based on pretrained I-JEPA [2] model that can extract rich visual representations from the image ($X_\mathcal{I}$). As shown in Figure 2, this auxiliary loss is implemented by predicting the visual features from the LLM backbone ($\mathcal{F}$) and matching them with the auxiliary representation using MSE loss as follows,

$$\begin{aligned}&\mathcal{L}_{visualLoss}(\theta_F,\theta_M,\theta_G) = \\ &\|\mathcal{A}(X_\mathcal{I}) - \mathcal{F}\Big(\mathcal{M}\big(\mathcal{G}(X_\mathcal{I})\big) \oplus \mathcal{T}\Big) \odot \texttt{vMask}\|^2\end{aligned} \tag{1}$$

where `vMask` is the binary mask that only retains the visual features in the output and masks out the textual features. The term $\mathcal{M}\big(\mathcal{G}(X_\mathcal{I})$ represents the visual tokens projected into LLM space and $\oplus\mathcal{T}$ represents the concatenation of textual tokens. The auxiliary loss is combined with the auto-regressive objective for LLM to train the overall MLLM using $\mathcal{L}_{tot}$,

$$\mathcal{L}_{tot} = \mathcal{L}_{ntp} + \beta \cdot \mathcal{L}_{visualLoss} \tag{2}$$

6.2. BlankTokens Formulation

Formally, we implement BlankTokens (`BlankInputsPartial` from Algorithm 1) as follows: say T_{in} are input text tokens, b_{id} is one of the reserved tokens from LLM vocabulary that we designate as blank token id, and $\mathcal{M}$ is the mask vector determining whether we want to blank an input token, then

$$T_{inBlank} = \begin{cases} T_{in} & \text{if } \mathcal{M} \text{ is True} \\ b_{Id} & \text{otherwise} \end{cases} \tag{3}$$

Our specific implementation of this strategy targets the initial tokens during response generation. We consistently blank out the first `N` text tokens of the input sequence provided to the model during training. This deliberate masking at the beginning of the sequence prevents the model from relying on leading textual cues to initiate its response. Instead, the model is forced to formulate its initial coherent thought and begin the generation process based primarily on the processed visual information. This targeted intervention aims to cultivate stronger visual grounding precisely at the critical starting point of generation, encouraging the model's subsequent output to remain more faithful to the visual context. Furthermore, beyond these initial tokens, we randomly blank out a fraction (about 20%) of the subsequent input tokens to discourage excessive reliance on language context and ensure the model continues to refer back to visual signals throughout its response generation.

6.3. Synthetic Data Generation

The generation process involves programmatically creating visual scenes paired with corresponding questions and answers. We randomly sample object instances from a large public image collection [7], and place a small number of these objects onto distinct locations within an $N \times N$ grid background. For each generated scene, we automatically formulate basic questions focusing on visual understanding (for example *What are the objects in the image?*), relative object direction (for example *In which direction is ... ?*) and simple distance (e.g. *What is the distance ... ?*). The ground truth answers are derived directly from the programmed spatial layout. Figure 4 shows examples of different grid layouts. The grids vary in background colors and grid size granularity, such as 4×4 or 8×8. We formulate 4 types of queries classified as: `Describe`, `Directional`, `Distance`, `Location`. Figure 5 shows examples of such queries.

Crucially, these generated questions are designed so they can typically only be answered correctly by accurately parsing the visual content of the grid and reasoning about the relative placement of the objects depicted. They inherently resist solutions based purely on language priors or statistical shortcuts. By integrating this synthetic data into our training, we complement our other innovations (VisualLoss, BlankTokens) by providing the model with explicit, targeted tasks that directly exercise and thereby sharpen its visual reasoning and modality alignment capabilities.

6.4. Training Details

Model Architecture. Our model architecture is closely based on the LLaVA framework [9]. We employ Llama 3.1 8B [5] as the LLM backbone. For visual feature extraction, we utilize the pre-trained CLIP backbone [10].

Figure 4. Examples of synthetic visual samples with objects placed on $N \times N$ grid with different backgrounds.

DESCRIBE:

What is shown in the image?

Answer: The image shows a canvas with 4x4 units.
The canvas is silver in color.
The canvas has 5 objects on it.
The objects are: Mobile phone at (3, 0), Bust at (0, 3), Hat at (0, 2), Common sunflower at (1, 1), Woodpecker at (3, 1).

DIRECTIONAL:

Here is an image of size 4x4 containing few objects. (0, 0) is the top-left corner and (3, 3) is the bottom-right corner. The first coordinate is the row and the second coordinate is the column. Provide answer to the following question by first elaborating the reasoning behind your answer in a step-by-step manner, followed by the answer in a word or phrase. Please format the final response as follows -- 'Reason: <reasoning behind the answer>' 'Answer: <single word/phrase answer to the question>'.
Question: What is the direction of the Mobile phone from the Common sunflower?

Reason: The Mobile phone is at (3, 0), and the Common sunflower at (1, 1).
Answer: left-down.

DISTANCE:

Here is an image of size 4x4 containing few objects. (0, 0) is the top-left corner and (3, 3) is the bottom-right corner. The first coordinate is the row and the second coordinate is the column. Provide answer to the following question by first elaborating the reasoning behind your answer in a step-by-step manner, followed by the answer in a word or phrase. Please format the final response as follows -- 'Reason: <reasoning behind the answer>' 'Answer: <single word/phrase answer to the question>'.
Question: Calculate the distance from the Hat to the Woodpecker.

Reason: The Hat is at (0, 2), and the Woodpecker at (3, 1).
Answer: 3.16.

LOCATION:

Here is an image of size 4x4 containing few objects. (0, 0) is the top-left corner and (3, 3) is the bottom-right corner. The first coordinate is the row and the second coordinate is the column. Provide answer to the following question by first elaborating the reasoning behind your answer in a step-by-step manner, followed by the answer in a word or phrase. Please format the final response as follows -- 'Reason: <reasoning behind the answer>' 'Answer: <single word/phrase answer to the question>'.
Question: Where can I find the Bust?

Reason: The Bust is at (0, 3).
Answer: (0,3).

Figure 5. Examples of spatial queries of type `Describe`, `Directional`, `Distance`, `Location` generated for a visual sample.

Training Setup. Our training follows a three-stage procedure similar to LLaVA. *Stage 1 (Feature Alignment):* We first train only the MLP connector module, keeping both the vision encoder and the LLM backbone frozen. We do not apply VISUALLOSS and BLANKTOKENS in this stage. *Stage 2 (End-to-End Continued Pretraining):* Subsequently, we train the MLP connector, LLM backbone and the image encoder jointly using a mixture of multimodal pretraining dataset. In this stage, we apply VISUALLOSS and BLANKTOKENS to encourage LLM to build a rich visual representation. *Stage 3 (End-to-End Instruction Fine-tuning):* Finally, we fine-tune the MLP connector, LLM backbone and the image encoder to answer questions based on visual and text information.

Technique Implementation & Data Mixture. When incorporating our auxiliary visual loss during training, we set the loss weighting coefficient $\beta = 0.5$. For our input masking technique aimed at reducing language prior reliance, we consistently blank out the first $N = 5$ text tokens as well as randomly selected 20% tokens in the input sequences. During the Stage 3 fine-tuning phase, our data mixture consists of 75% standard multimodal instruction data and 25% of our targeted synthetic dataset described in Section 3.4.

6.5. Related Work

Our approach relates to but diverges from several lines of prior work aimed at improving MLLM visual grounding. We highlight two key areas:

Multiple Image Encoders. A substantial body of work has explored enhancing visual input by feeding representations from multiple distinct image encoders into the LLM backbone [14, 17, 18]. While providing more diverse visual information can be marginally helpful, this strategy often struggles to significantly improve the LLM's core visual un-

derstanding. This limitation arises partly because the standard next-token prediction loss used in LLM training is not inherently conducive to integrating rich, fine-grained visual information deeply into the LLM's latent space.

Auxiliary Visual Losses. Another relevant direction involves cross-modal attention by incorporating auxiliary losses specifically designed to promote region-level understanding or grounding during training [13, 22, 23]. These methods aim to provide more direct supervision for visual interpretation (GlaMM [13] offers a detailed analysis related to auxiliary visual input). However, a common downside of such approaches is the potential need for extra visual signals or annotations (e.g., object locations, regional descriptions) to compute these losses. In contrast, our work focuses on enhancing the LLM's internal visual representation through an auxiliary loss applied in a self-supervised manner, leveraging the structure learned by powerful image encoders without requiring explicit region-level annotations for the auxiliary task itself.

6.6. Next-token Prediction Visualization

We make the following observations regarding the next-token prediction accuracy shown in the figure:

- **Higher-quality captions:** We observe that the first few tokens are consistently more accurate for our model compared to the baseline. This demonstrates that our BLANKTOKENS intervention enhances the model's ability to compose text grounded in the visual content from the outset.
- **Higher confidence prediction of visually relevant text tokens:** We observe that tokens pertaining to visual concepts (such as objects, people, orientation, shape, and color) are predicted with higher confidence by the model trained with our innovations.
- **Focus on enriching cross-modal alignment:** The baseline model achieves its highest prediction accuracy primarily for language-centric tokens like `contains`, `in the`, and `side`. This further supports the conclusion that our approach's primary impact is on improving the generation of visually grounded text requiring reference to visual input and reasoning.

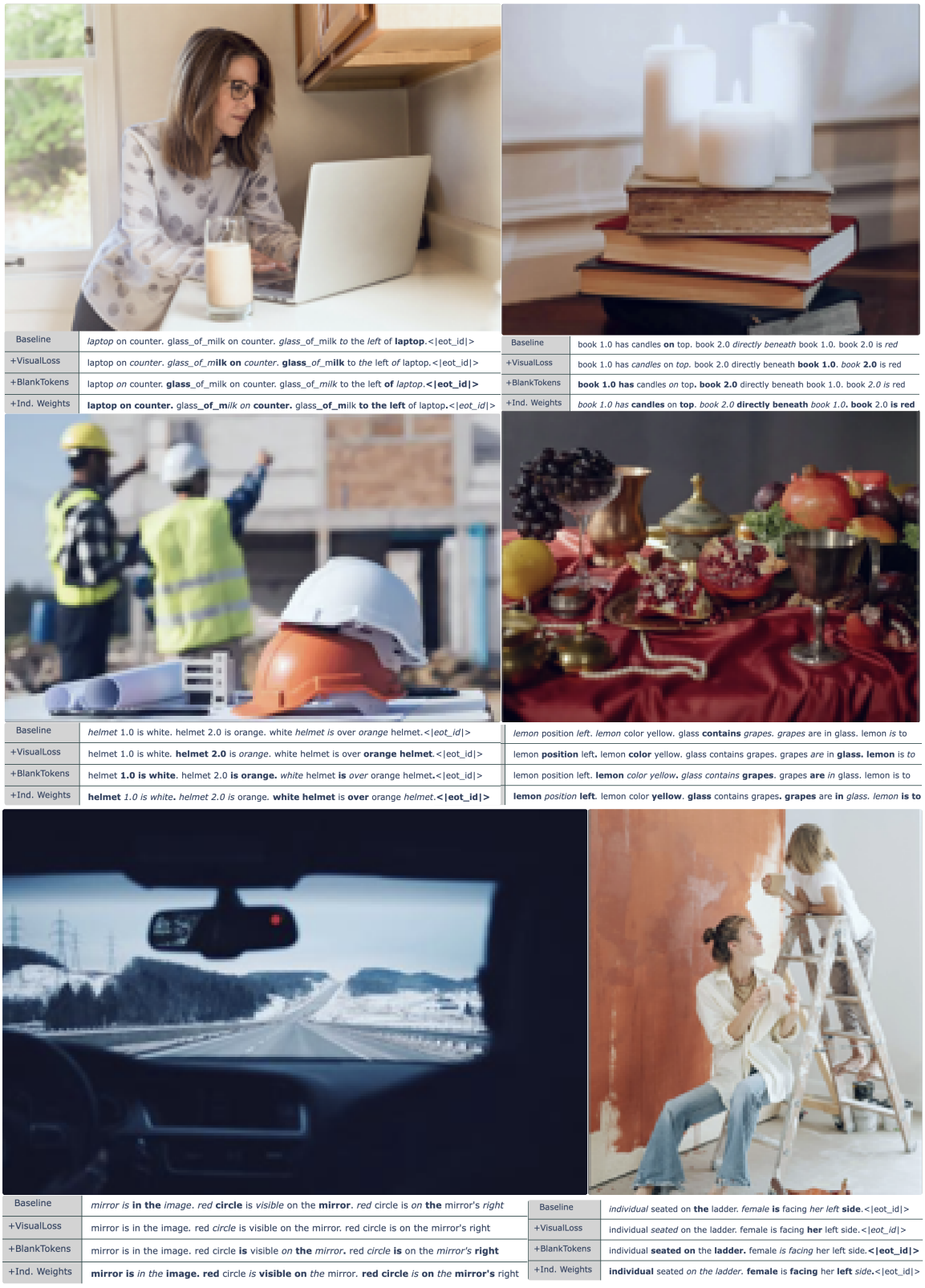


Figure 6. This figure visualizes the next-token prediction loss on a per-token basis for our proposed approaches. The visualization highlights improvements in the model's ability to predict tokens corresponding to visual content. In the figure, **bold** text represents the lowest cross entropy loss when comparing across the model variants.