Too Late to Recall: The Two-Hop Problem in Multimodal Knowledge Retrieval

Constantin Venhoff
University of Oxford

Ashkan Khakzar
University of Oxford

Sonia Joseph
McGill University / Meta

Philip Torr
University of Oxford

Neel Nanda

Abstract

LLaVA-style Vision-Language Models (VLMs) have demonstrated impressive capabilities, but struggle with factual recall tasks compared to their underlying language model (LM). While previous work attributes this to insufficient computational depth after visual processing, we provide an alternative explanation: the distributed representations of visual information across visual tokens in early layers bypasses the factual recall mechanism that resides in the early-layer MLPs of the LM backbone. The performance gap therefore stems from the architectural design of VLMs, rather than insufficient computational capacity. Using linear probes, we show that dedicated linear representations of visual information only emerge in the middle-to-late layers of VLMs. As a result, factual recall in VLMs becomes a "two-hop" challenge, where factual recall precedes visual processing, but the visual processing finishes too late in the model. Through comparative analysis, we demonstrate that successful factual recall depends on the speed of the first processing "hop." To further support our hypothesis, we patch early-layer MLP outputs from the LM backbone into the corresponding VLM layers, significantly improving factual recall performance. This suggests that the absence of properly aligned token embeddings in early layers is a key factor in factual recall degradation. Finally, we introduce a benchmark to systematically evaluate factual recall accuracy and knowledge hallucination in multimodal settings. Our findings highlight a fundamental architectural limitation in current VLMs and pave the way for designing models that better integrate visual and linguistic information for reliable factual reasoning.

1. Introduction

Vision-Language Models (VLMs) achieve strong multimodal reasoning by integrating vision transformers (ViTs) with large language models (LLMs) via adapter mechanisms [6, 10–12, 18, 19]. These adapters project visual representations into the LLM's token embedding space, enabling text-based reasoning over visual inputs. However, recent work has found that these visual projections are misaligned with the pretrained token embeddings of the LLM [14], while separate studies have identified significant factual recall degradation in VLMs [7]. These two issues, however, have not yet been directly connected. We hypothesize that adapter misalignment is the root cause of factual recall degradation in VLMs. Specifically, we argue that misaligned visual projections prevent early LLM layers—where factual recall mechanisms are known to reside [5, 8, 15]—from engaging effectively. This creates a two-hop problem, where the model must first resolve visual inputs into structured entity representations before retrieving associated knowledge. Prior research shows that LLMs already struggle with such two-hop reasoning tasks in purely text-based settings [23], underscoring the challenge in VLMs. To investigate, we perform linear probing and patching experiments. We show that visual entity representations emerge too late in the LLM forward pass, bypassing early-layer factual recall mechanisms. By patching early-layer MLP outputs from the unimodal LLM into VLM layers, we significantly recover factual recall, providing strong causal evidence for our hypothesis. Furthermore, to validate our findings at scale, we introduce a large-scale benchmark designed to systematically measure factual recall in VLMs versus the base version of their language model backbone. Using this benchmark, we evaluate multiple VLMs—including LLaVA-1.5-7B, LLaVA-1.5-13B [12], LLaVA-MORE [6], GPT-4o, and Claude Opus, confirming that factual recall degradation is a consistent and widespread issue. Our findings highlight a fundamental architectural limitation in VLMs and underscore the need for improved adapter alignment to enhance factual reasoning.

2. The Token Space Misalignment

To investigate the alignment between projected visual embeddings and pretrained textual token embeddings in Vision

Language Models (VLMs), we conduct an empirical analysis of visual projector outputs. Prior work suggests that effective factual recall in LLMs relies on structured token embeddings in early layers [5, 8, 15]. If visual embeddings are systematically misaligned with these pretrained token representations, they may fail to activate the LLM's factual recall mechanisms, contributing to degraded multimodal knowledge retrieval.

Experimental Setup: We evaluate three VLMs—LLaVA-1.5-7B, LLaVA-1.5-13B, and LLaVA-MORE, by extracting visual projector outputs from 1,000 randomly selected ImageNet images. Each image produces 575 visual token activations, yielding a total of 575,000 activations per model. To assess alignment, we randomly sample 10,000 visual token embeddings per model and compute their cosine similarity with textual token embeddings from the LLM backbone.

Results and Analysis: Figure 1 presents the distribution of cosine similarity scores between visual projector outputs and textual embeddings. The results indicate a pronounced misalignment across all evaluated models, with cosine similarity scores tightly concentrated near zero. This suggests that visual tokens predominantly occupy an embedding subspace orthogonal to pretrained textual representations.

These findings provide direct empirical support for our hypothesis that visual embeddings fail to integrate naturally into the LLM backbone's structured token space. As a result, early-layer factual recall mechanisms remain disengaged when processing visual inputs, reinforcing the idea that factual recall degradation in VLMs stems from fundamental adapter misalignment rather than insufficient computational depth alone.

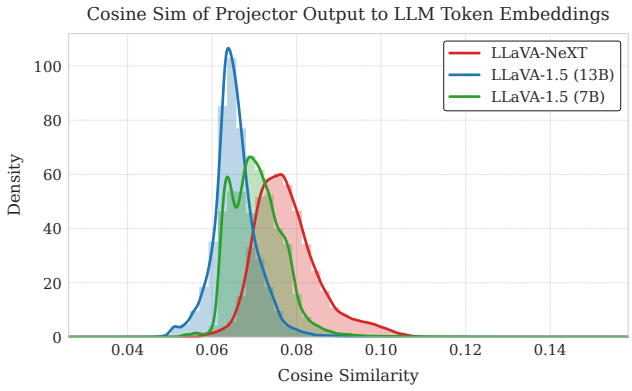

Figure 1. Distribution of absolute cosine similarity between visual projector outputs and text embeddings for LLaVA-1.5-7B, LLaVA-1.5-13B, and LLaVA-MORE. All models exhibit similar distributions, with means ranging from approximately 0.06 to 0.08. The distributions are characterized by their sharp, concentrated profiles with high density around their respective means. This concentrated distribution pattern strongly indicates that visual token embeddings consistently occupy subspaces nearly orthogonal to the pretrained text token embedding space across all evaluated models.

3. Benchmarking Factual Recall in Vision-Language Models

To systematically evaluate factual recall degradation in Vision-Language Models (VLMs), we introduce a benchmark that directly compares their performance against their corresponding language-only backbone models. This benchmark isolates the impact of factual recall capabilities by ensuring that equivalent input information is accessible to both the VLM and the language model backbone.

3.1. Benchmark Design

Our benchmark consists of 1,000 factual recall questions designed to assess multimodal knowledge retrieval. We sample images fom the ImageNet dataset and use GPT-4o to generate entity-specific factual questions (e.g., "Who invented the entity shown in the image?"). GPT-4o was selected as it performed best on the WildHallucinations benchmark, which rigorously evaluates factual accuracy in language models [25].

During evaluation, each VLM is first prompted to identify the main entity depicted in each image. Samples where the entity is misidentified are excluded to prevent compounding factual retrieval with recognition errors. For correctly identified entities, we assess factual recall accuracy by comparing the VLM's answer to that of its language-only backbone, ensuring a controlled comparison.

3.2. Benchmark Results

We evaluate a range of VLMs, including LLaVA-1.5-7B (Llama-2-7B-it), LLaVA-1.5-13B (Llama-2-13B-it), and LLaVA-MORE (Llama-3.1-8B-it), alongside frontier models such as GPT-4o and Claude Opus. As shown in Figure 2, all VLMs exhibit significant factual recall degradation relative to their unimodal counterparts. This result strongly supports our hypothesis that misaligned visual embeddings disrupt access to pretrained factual knowledge in early layers, leading to reduced factual retrieval performance.

These findings highlight a fundamental limitation in current VLM architectures and underscore the need for improved adapter alignment to better integrate visual representations into the language model's factual reasoning processes.

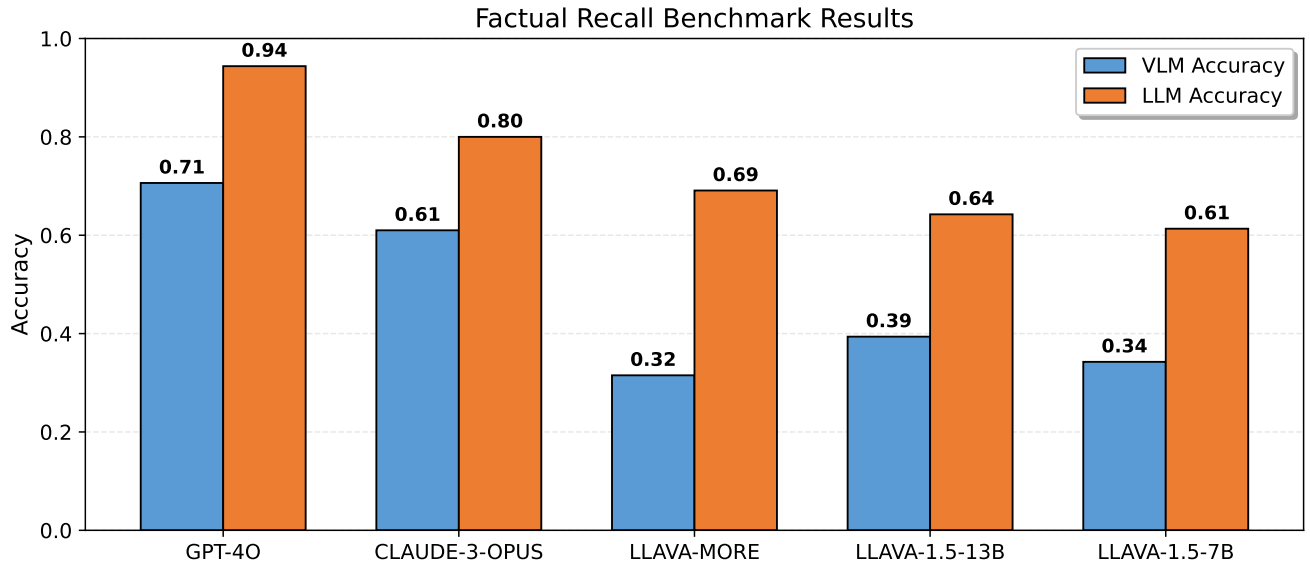

Figure 2. Comparison of factual recall accuracy between multimodal Vision-Language Models (VLMs) and the base-version of their LLM backbones. All tested models show substantial performance degradation in the multimodal setting compared to their unimodal counterparts. State-of-the-art frontier models (GPT-4o and Claude Opus) demonstrate a somewhat reduced but still significant performance gap compared to open-source models (LLaVA variants), suggesting this phenomenon transcends specific architectures and exists even in the most advanced models. This consistent pattern provides compelling evidence that the factual recall degradation represents a fundamental challenge in multimodal language processing rather than an implementation-specific limitation.

4. Closing Performance Gap through Activation Patching

To further investigate the role of embedding misalignment in factual recall degradation, we identify the specific MLPs in the language model backbone that are responsible for producing enriched factual token representations and test whether restoring these representations in VLMs can recover factual recall performance. Our approach consists of two complementary stages: (*i*) attributing factual recall to specific MLP layers in the unimodal LLM backbone, and (*ii*) performing cross-model patching by injecting enriched early-layer outputs from the unimodal model into corresponding layers of the VLM.

4.1. Attribution Patching in the Language Model Backbone

To determine which MLPs in the language model contribute most to factual recall, we employ an attribution patching methodology inspired by Nanda and Meng et al.. First we sample 100 examples from the benchmark dataset for each language model backbone, which were answered correctly by model. Then we use the following steps to compute the attriution scores:

1. **Corrupting Entity Representations**: We first generate a corrupted input where the entity token embeddings $\mathbf{h}_e$ are replaced with the mean token embedding $\mathbf{h}^{\text{mean}}$ computed across the whole attribution datasets, which degraded the model's ability to recognize and recall facts about the entity:

$$\mathbf{h}_e^{\text{corrupt}} = \mathbf{h}^{\text{mean}}.$$

2. **Computing KL Divergence Gradients**: We run the model on the corrupted input and compute the KL divergence $D_{\text{KL}}(P_{\text{corrupt}} \| P_{\text{clean}})$ between the corrupted and clean output distributions, where the clean distribution is obtained from running the non-corrupted example through the language model. The gradient of this divergence with respect to the MLP outputs provides an estimate of the causal importance of each layer.
3. **Caching Clean and Corrupted MLP Activations**: We store the MLP outputs $\mathbf{h}_\ell^{\text{clean}}$ from the clean run and $\mathbf{h}_\ell^{\text{corrupt}}$ from the corrupted run for each layer ℓ.
4. **Computing Attribution Scores**: The absolute change in KL divergence when restoring clean MLP outputs defines the attribution score:

$$\mathcal{A}_\ell = \left| (\mathbf{h}_\ell^{\text{clean}} - \mathbf{h}_\ell^{\text{corrupt}}) \cdot \nabla_{\mathbf{h}_\ell^{\text{corrupt}}} D_{\text{KL}} \right|.$$

Layers with the highest attribution scores are identified as the primary contributors to enriched factual recall representations.

Figure 3 shows the attribution scores averaged over the 100 correct factual recall examples from the benchmark dataset,

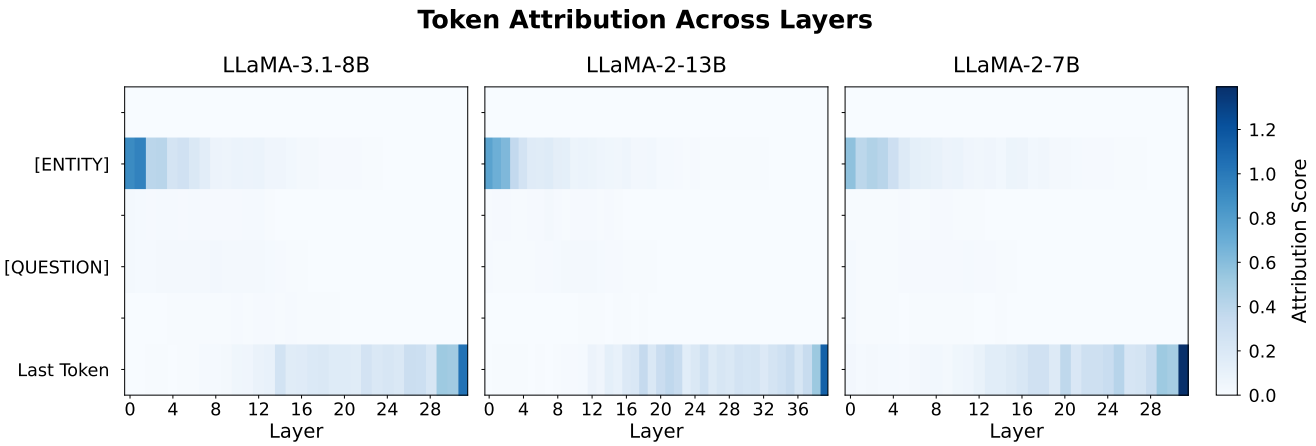

Figure 3. Attribution scores of each MLP layer in the Llama-3.1-8B-it, Llama-2-13B, and Llama-2-7B language model backbones, quantifying their contributions to enriched entity representation during factual recall tasks. Token positions for the entity, question, last token, and intermediate tokens are aggregated into the averaged score for readability. Higher values indicate greater causal relevance. The figure reveals a pronounced concentration of entity token relevance in the early layers across all three models. This consistent early-layer specialization across different model scales provides a mechanistic explanation for why visual token misalignment at these stages could significantly impair factual recall capabilities in multimodal contexts.

aggregated across different token positions (entity, question, and last token positions), averaged into a single position for readability. We find that the MLP outputs at the entity token positions are causally most relevant (besides trivially the last token/layer position).

Cross-Model Patching Having identified the critical MLP layers responsible for factual recall, we conduct a cross-model patching intervention to restore enriched representations in VLMs. Specifically, we inject MLP outputs from the layers 0-5 at the entity token positions of the language model backbone into corresponding layers of the VLM during factual recall tasks. Since VLMs lack explicit token positions for visual entities, we employ a greedy matching strategy to select token positions for patching based that yield minimal KL divergence between the patched VLMs predictive distribution, and the language model backbone's one. We apply this intervention to all evaluation examples where the VLM initially failed factual recall despite correctly recognizing the entity visually, while the language model succeeded.

4.2. Patching Results

Figure 4 illustrates the impact of patching early-layer representations from the unimodal backbone into the VLM. This intervention recovers almost 40% of the factual recall accuracy difference between the VLM and its language-only counterpart. Given that enriched subject representations are just one part of the factual recall mechanism described by Chughtai et al., an almost 40% performance recovery support our hypothesis that factual recall degradation in multimodal models is primarily due to the absence of properly

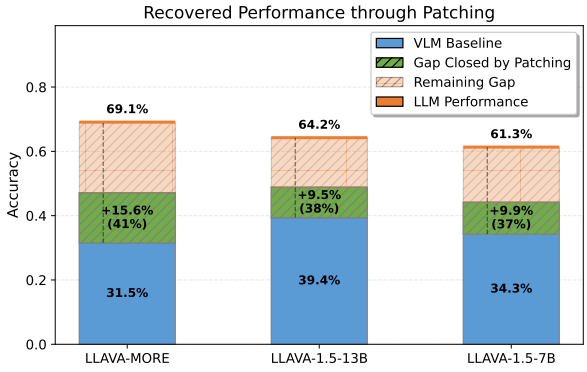

Figure 4. Factual recall performance recovery for LLaVA-1.5-7B, LLaVA-1.5-13B, and LLaVA-MORE models when enriched entity representations from corresponding language model backbones are patched into early layers. Each model's performance is represented in a stratified visualization showing the baseline VLM accuracy (blue), the performance recovery achieved through early-layer patching (green), and the target LLM backbone performance (orange). The significant recovery observed across all tested architectures (almost 40% of the original performance gap) provides compelling evidence that the misalignment of visual embeddings in early layers directly contributes to factual recall degradation in multimodal contexts.

aligned and enriched entity representations in early LLM layers.

Overall, this experiment provides strong causal evidence that embedding misalignment in early layers is a fundamental driver of factual recall degradation in VLMs, reinforcing the need for better-aligned adapter mechanisms to improve

multimodal factual reasoning.

5. Probing for the Emergence of Visual Entity Representations

While previous experiments demonstrated a clear misalignment between visual projector outputs and textual embeddings, an alternative explanation could suggest that multimodal models quickly infer a linear representation of visual entities in early layers, potentially mitigating the adverse effects of embedding misalignment. In this section, we empirically investigate at which network layers visual entities become linearly represented in VLMs and how this relates to factual recall performance.

5.1. Linear Probing for Visual Entity Representations

To systematically assess when visual entity representations emerge within the language backbone of VLMs, we employ a probing methodology. Specifically, we train layer-wise linear classifiers (probes) on the residual stream outputs of each transformer layer to predict visual entities depicted in input images.

Experimental Setup: We use CIFAR-100 as our evaluation dataset due to its controlled set of 100 classes and multiple images per entity, ensuring reliable estimation of representational capacity. We randomly select 5,000 CIFAR-100 images and construct diverse natural language prompts instructing the model to describe the image (e.g., "Identify the object shown," "What do you see in this image?"). At each transformer layer, we train an independent linear probe on the extracted and averaged residual-stream representations at the text token positions, to predict the correct entity label. We use a 20%/80% train-test split to evaluate the probes.

Results and Analysis: Figure 5 depicts the accuracy of the linear probes across layers for LLaVA-1.5-7B, LLaVA-1.5-13B and LLaVA-MORE. The results reveal a clear trend: linear representations capable of reliably encoding visual entity information do not emerge until the middle-to-late layers. Prior to these middle layers, probe accuracy remains poor, indicating that early layers do not encode robust entity representations. These findings directly refute the notion that multimodal models infer linear visual representations in early layers, instead supporting the hypothesis that VLMs spend a substantial portion of their time processing visual inputs before reaching a stage where factual recall can be engaged. This aligns with our **two-hop hypothesis**: the model must first form a structured entity representation (first hop) before retrieving factual knowledge (second hop). However, since entity representations only emerge in deeper layers, they bypass the early-layer factual recall mechanisms, contributing to factual recall degradation.

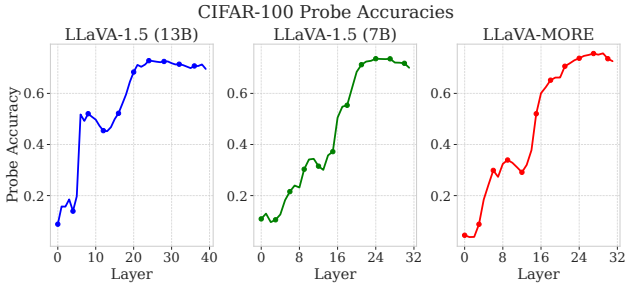

Figure 5. Accuracy of linear probes trained on residual-stream representations at each transformer layer of LLaVA-1.5-7B, LLaVA-1.5-13B and LLaVA-MORE measured on CIFAR-100 entity prediction. All three models exhibit a consistent pattern: probe accuracy remains poor in early layers and rises sharply between middle-to-late layers. This pronounced inflection point reveals the delayed emergence of linearly probeable visual entity representations, occurring well after the critical early-layer factual recall mechanisms identified in our attribution analysis.

5.2. Comparing Successful vs. Unsuccessful Factual Recall Cases

To further investigate the role of early-layer entity representations in factual recall, we compare cases where factual recall succeeds versus fails, hypothesizing that earlier emergence of entity representations correlates with higher factual recall accuracy, aligning with our two-hop hypotheses.

Experimental Setup: We generate a multimodal factual recall dataset using CIFAR-100 images, following the methodology in Section 3. For each image, we construct factual queries requiring both entity recognition and factual knowledge retrieval. We then separate the model's responses into two groups:

- **Successful factual recall:** Cases where the VLM correctly answers the factual query.
- **Unsuccessful factual recall:** Cases where the VLM fails to answer correctly despite recognizing the entity.

Each VLM answers around 1000 examples correctly. Given the smaller training set size, we do a 50%/50% test-train split, to ensure we have sufficient evaluation examples. For each group, we train linear probes similar to the previous probing setup in 5.1 and record the evaluation accuracy and the training loss.

Results and Analysis: Figure 6 shows a clear divergence in probe accuracy and training loss between the two groups in earlier layers, before converging onto the same accuracy and loss in later layers. These findings strongly support

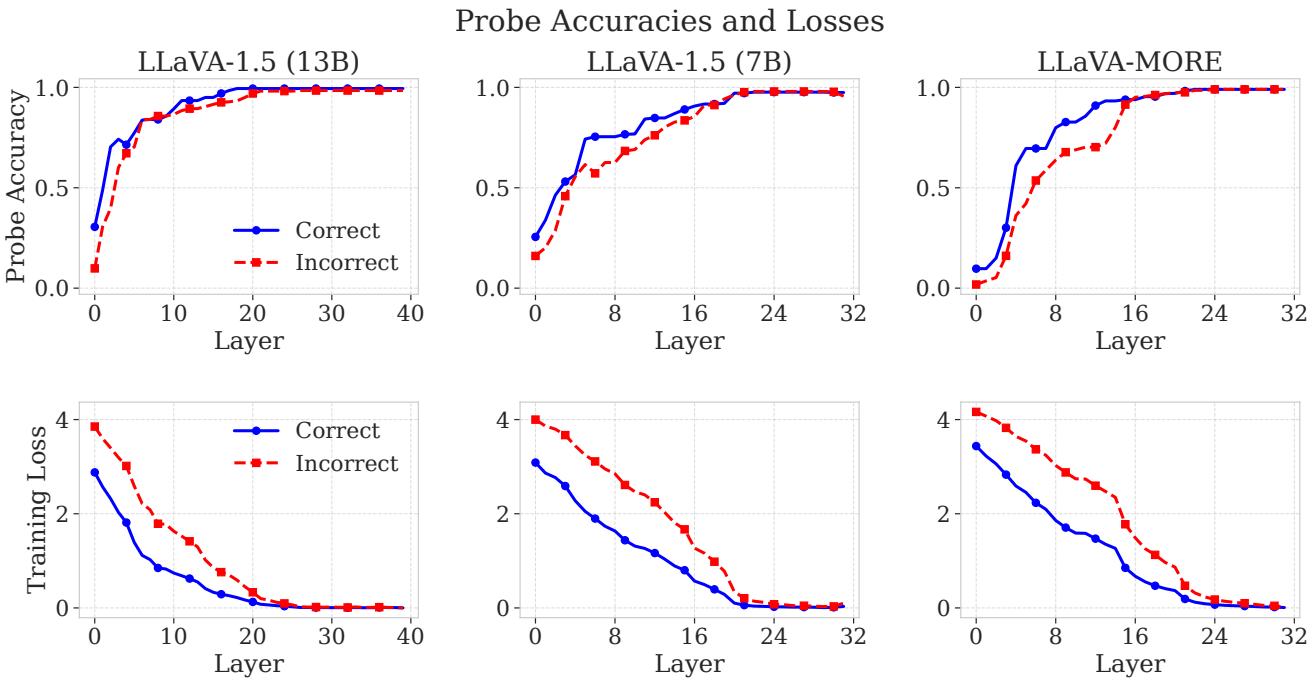

Figure 6. Comparison of probe accuracy and training loss across layers for correct versus incorrect factual recall examples in LLaVA-1.5-7B, LLaVA-1.5-13B and LLaVA-MORE. Probes trained on successful factual recall cases show higher accuracy in early layers and lower loss compared to unsuccessful cases, highlighting the importance of timely entity representation formation. This distinctive pattern demonstrates that when factual recall succeeds, visual entity information becomes identifiable through linear probing significantly earlier in the processing pipeline, while this emergence is delayed in unsuccessful cases. As processing progresses through deeper layers, the accuracy gap diminishes, suggesting that all examples eventually form adequate entity representations, but the early-layer advantage appears to be a key determinant of successful factual recall.

our two-hop hypothesis: factual recall success is highly dependent on the timely formation of entity representations. When entity information is encoded earlier, it can more effectively engage with the factual recall mechanisms that predominantly operate in the early layers of the language model backbone. The late emergence of visual entity representations in unsuccessful cases leads to misalignment with these mechanisms, resulting in factual recall failures and being crucially unrelated to remaining computational capacity.

Conclusion: This analysis provides strong empirical evidence that the temporal alignment between entity recognition and factual recall mechanisms is a key determinant of factual recall success in VLMs. The late emergence of visual entity representations prevents effective engagement with early-layer recall mechanisms, reinforcing embedding misalignment as a fundamental bottleneck in multimodal factual reasoning.

6. Related Work

In this section, we review prior research on Vision-Language Models (VLMs) and their integration of visual and textual representations. We first discuss approaches to aligning visual tokens with language model embeddings and their impact on multimodal reasoning. We then examine studies on factual recall in both unimodal and multimodal models, highlighting recent findings on factual recall degradation in VLMs. Finally, we contrast our work with existing research, emphasizing how our probing and patching experiments provide new insights into the mechanisms underlying multimodal factual retrieval.

6.1. Vision Language Models (VLMs)

Vision Language Models operate by projecting visual and textual information into an embedding space before processing the resulting tokens through an LLM, pretrained on text [11]. A critical factor for VLMs performance is properly mapping visual features into token space while preserving essential visual information (e.g., entity recognition) [14]. In the simplest architectural configuration, cross-modal interaction is achieved through image encoders coupled with a projector (typically implemented as a multilayer perceptron, MLP), which maps visual data into the embedding space of text tokens—specifically, into the text embedding manifold that the LLM was pre-trained to pro-

cess [4, 11, 13]. However, this approach does not always guarantee actual alignment between textual and visual token embeddings, potentially degrading VLM performance [10, 14]. To improve the alignement, researchers have proposed alternative methods to map in the embedding (see [1, 9, 20, 21]) Alternative methods enhance cross-modal interaction at deeper network layers, rather than relying on perfect vision-text token alignment, by incorporating cross-attention and feed-forward layers throughout the LLM architecture [2, 24]. While effective, these approaches significantly increase computational requirements and parameter counts.

6.2. Factual Recall in LLMs and VLMs

Large Language Models effectively process text sequences, generate coherent outputs, and respond to factual queries (e.g., "Who invented transistors?"). Recent studies have established that early network layers predominantly handle factual recall processes [5, 8, 15]. This localization of factual processing offers a potential explanation for LLMs' performance degradation when confronted with multi-hop reasoning tasks [3, 17, 22], requiring sequential question resolution. For instance, answering "Who is the mother of the inventor or transistors" necessitates first resolving "Who is the inventor of transistors?" before addressing the primary question, creating challenges for models with factual processing concentrated in early architectural layers. To address this challenge, [3] proposes that early network layers focus on resolving the "bridge entity" (e.g., the "inventor or transistors"), while the secondary reasoning step must be processed by later layers that may have diminished factual recall capabilities [5].

Factual recall degradation in VLMs: Recent work [7] is the first to report factual recall degradation in VLMs (in LLaVA 1.5), revealing through patching experiments that visual representations cease to be leveraged for entity recognition beyond certain layers (middle layers). They postulate that layers prior to this threshold are primarily occupied with visual token processing. Our work extends this understanding through complementary probing experiments that provide a more direct measurement of when visual representations become linearly recognizable within the LLM architecture. This approach differs from the entity patching experiments in [7], which serve as a proxy measurement for visual token utilization rather than offering direct insight into visual information processing. Furthermore, while [7] attributes factual recall challenges to insufficient remaining layers for factual processing after visual processing completes, our findings reveal a different mechanism: factual recall predominantly occurs in early layers, but visual representations aren't processed until later layers, causing these critical factual recall mechanisms to be bypassed entirely. Although [7] provides valuable initial insights, our experiments shed new light on the underlying mechanisms of this phenomenon.

7. Discussion and Conclusion

This work systematically investigates the root causes of factual recall degradation in Vision-Language Models (VLMs), identifying the misalignment between adapter outputs and language model embeddings as a key factor. Our findings reveal that visual adapter outputs remain systematically misaligned with the pretrained token embeddings of the language model backbone, preventing effective engagement with early-layer factual recall mechanisms. Using attribution patching, we demonstrate that factual recall primarily relies on early-layer MLPs in the LLM, and restoring their outputs in VLMs significantly improves recall performance. Furthermore, probing experiments show that visual entity representations emerge too late in the network, only after the middle-to-late layers, bypassing the critical early-layer recall mechanisms. Comparing successful and unsuccessful recall cases further confirms that earlier formation of entity representations strongly correlates with higher recall accuracy, supporting our two-hop hypothesis: VLMs must first resolve visual entities (first hop) before retrieving factual knowledge (second hop), but misalignment disrupts the first step at a crucial stage.

These findings highlight fundamental architectural limitations in current VLMs and suggest several avenues for improvement. First, adapter mechanisms require better alignment techniques to integrate visual tokens meaningfully into the LLM's embedding space. Second, architectural modifications should encourage earlier formation of visual entity representations, ensuring they can effectively engage factual recall mechanisms. Finally, hybrid approaches that dynamically route visual information into early factual recall layers may help mitigate these issues.

While our study focuses on LLaVA-style models, future work should investigate whether these limitations persist across different architectures. Additionally, evaluating larger-scale VLMs may provide insights into whether model size mitigates or exacerbates these issues. Further research into alternative probing techniques and controlled interventions could refine our understanding of multimodal representation alignment.

Overall, this study provides strong empirical evidence that embedding misalignment fundamentally limits factual recall in VLMs. Addressing this issue is crucial for improving multimodal factual reasoning and reducing hallucinations in vision-language tasks. Future research should prioritize developing alignment-aware architectures to bridge the gap between vision and language in large-scale multimodal models.

References

[1] Sravanti Addepalli, Ashish Ramayee Asokan, Lakshay Sharma, and R Venkatesh Babu. Leveraging vision-language models for improving domain generalization in image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23922–23932, 2024. 7

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 7

[3] Eden Biran, Daniela Gottesman, Sohee Yang, Mor Geva, and Amir Globerson. Hopping too late: Exploring the limitations of large language models on multi-hop queries. *arXiv preprint arXiv:2406.12775*, 2024. 7

[4] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024. 7

[5] Bilal Chughtai, Alan Cooney, and Neel Nanda. Summing up the facts: Additive mechanisms behind factual recall in llms. *arXiv preprint arXiv:2402.07321*, 2024. 1, 2, 4, 7

[6] Federico Cocchi, Nicholas Moratelli, Davide Caffagni, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. LLaVA-MORE: Enhancing Visual Instruction Tuning with LLaMA 3.1, 2024. 1

[7] Ido Cohen, Daniela Gottesman, Mor Geva, and Raja Giryes. Performance gap in entity knowledge extraction across modalities in vision language models. *arXiv preprint arXiv:2412.14133*, 2024. 1, 7

[8] Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. *arXiv preprint arXiv:2304.14767*, 2023. 1, 2, 7

[9] Jinhao Li, Haopeng Li, Sarah Erfani, Lei Feng, James Bailey, and Feng Liu. Visual-text cross alignment: Refining the similarity score in vision-language models. *arXiv preprint arXiv:2406.02915*, 2024. 7

[10] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26689–26699, 2024. 1, 7

[11] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 6, 7

[12] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 1

[13] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llavanext: Improved reasoning, ocr, and world knowledge, 2024. 7

[14] Ahmed Masry, Juan A Rodriguez, Tianyu Zhang, Suyuchen Wang, Chao Wang, Aarash Feizi, Akshay Kalkunte Suresh, Abhay Puri, Xiangru Jian, Pierre-André Noël, et al. Alignvlm: Bridging vision and language latent spaces for multimodal understanding. *arXiv preprint arXiv:2502.01341*, 2025. 1, 6, 7

[15] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022. 1, 2, 3, 7

[16] Neel Nanda. Attribution patching: Activation patching at industrial scale. *URL: https://www. neelnanda. io/mechanistic-interpretability/attribution-patching*, 2023. 3

[17] Mansi Sakarvadia, Aswathy Ajith, Arham Khan, Daniel Grzenda, Nathaniel Hudson, André Bauer, Kyle Chard, and Ian Foster. Memory injections: Correcting multi-hop reasoning failures during inference in transformer-based language models. *arXiv preprint arXiv:2309.05605*, 2023. 7

[18] Andreas Steiner, André Susano Pinto, Michael Tschannen, Daniel Keysers, Xiao Wang, Yonatan Bitton, Alexey Gritsenko, Matthias Minderer, Anthony Sherbondy, Shangbang Long, Siyang Qin, Reeve Ingle, Emanuele Bugliarello, Sahar Kazemzadeh, Thomas Mesnard, Ibrahim Alabdulmohsin, Lucas Beyer, and Xiaohua Zhai. Paligemma 2: A family of versatile vlms for transfer, 2024. 1

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

[20] Uddeshya Upadhyay, Shyamgopal Karthik, Massimiliano Mancini, and Zeynep Akata. Probvlm: Probabilistic adapter for frozen vison-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1899–1910, 2023. 7

[21] Dawei Yan, Pengcheng Li, Yang Li, Hao Chen, Qingguo Chen, Weihua Luo, Wei Dong, Qingsen Yan, Haokui Zhang, and Chunhua Shen. Tg-llava: Text guided llava via learnable latent embeddings. *arXiv preprint arXiv:2409.09564*, 2024. 7

[22] Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. Do large language models latently perform multi-hop reasoning? *arXiv preprint arXiv:2402.16837*, 2024. 7

[23] Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. Do large language models latently perform multi-hop reasoning? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10210–10229, Bangkok, Thailand, 2024. Association for Computational Linguistics. 1

[24] Zhuolin Yang, Wei Ping, Zihan Liu, Vijay Korthikanti, Weili Nie, De-An Huang, Linxi Fan, Zhiding Yu, Shiyi Lan, Bo Li, et al. Re-vilm: Retrieval-augmented visual language model for zero and few-shot image captioning. *arXiv preprint arXiv:2302.04858*, 2023. 7

[25] Wenting Zhao, Tanya Goyal, Yu Ying Chiu, Liwei Jiang, Benjamin Newman, Abhilasha Ravichander, Khy-

athi Chandu, Ronan Le Bras, Claire Cardie, Yuntian Deng, et al. Wildhallucinations: Evaluating long-form factuality in llms with real-world entity queries. *arXiv preprint arXiv:2407.17468*, 2024. 2