

LaMI: Augmenting Large Language Models via Late Multi-Image Fusion

Guy Yariv¹ Idan Schwartz^{2*} Yossi Adi^{1*} Sagie Benaim^{1*}

¹The Hebrew University of Jerusalem ²Bar-Ilan University
guy.yariv@mail.huji.ac.il

Abstract

Commonsense reasoning often requires both textual and visual knowledge, yet Large Language Models (LLMs) trained solely on text lack visual grounding (e.g., “what color is an emperor penguin’s belly?”). Visual Language Models (VLMs) perform better on visually grounded tasks but face two limitations: (i) often reduced performance on text-only commonsense reasoning compared to text-trained LLMs, and (ii) adapting newly released LLMs to vision input typically requires costly multi-modal training. An alternative augments LLMs with test-time visual signals, improving visual commonsense without harming textual reasoning, but prior designs often rely on early fusion and a single image, which can be suboptimal. We propose a *late multi-image fusion* method: multiple images are generated from the text prompt with a lightweight parallel sampling, and their prediction probabilities are combined with those of a text-only LLM through a late-fusion layer that integrates projected visual features just before the final prediction. Across visual commonsense and NLP benchmarks, our method significantly outperforms augmented LLMs on visual reasoning, matches VLMs on vision-based tasks, and, when applied to strong LLMs such as LLaMA 3, also improves NLP performance while adding only modest test-time overhead.

1 Introduction

Large Language Models (LLMs) advance a wide range of language tasks (Devlin et al., 2019; Radford et al., 2019; Zhang et al., 2022b; Team et al., 2024; Touvron et al., 2023), but training on text alone leaves them weak in visual commonsense. Vision–Language Models (VLMs) jointly train on images and text (Alayrac et al., 2022; Liu et al., 2023b,a; Li et al., 2023a; Dai et al., 2023; Cha et al., 2024), improving visually grounded abilities

Prompt: “Answer List: brown, black, white, yellow, green, gray, red, orange, blue, silver, and pink. What is the color of an emperor penguin’s belly? Please select the most possible answer from the above list. Please answer in one word.”

Llama3-8B-Instruct: “Yellow.” ✗
Ours (Llama3-8B-Instruct): “White.” ✓



Figure 1: LAMI at inference. While a text-only LLM fails on penguin belly color, LAMI generates diverse visual evidence that is late-fused and aggregated to correct the prediction and derive a grounded output.

such as VQA (Zhang et al., 2022a; Xia et al., 2023; Li et al., 2023b; Jin et al., 2024). However, they require heavy multimodal training and can reduce non-visual language performance, making rapid adaptation to new LLMs costly. The question is how to add robust visual knowledge to text-only models efficiently.

Visually-augmented LLMs (VaLMs) inject visual signals into pretrained LLMs without full multimodal retraining (Wang et al., 2023; Guo et al., 2023; Zhang et al., 2022a; Cui et al., 2024; Tan and Bansal, 2020), improving visual commonsense and sometimes even non-visual tasks (Zhang et al., 2024; Lu et al., 2022; Tang et al., 2023; Zhang et al., 2022a; Yang et al., 2022; Huang et al., 2023). Yet many methods fuse modalities early and rely on a single image, which can disturb LLM behavior and introduce noise and biases.

We propose a VaLM-style approach that addresses these issues through two key components. First, we introduce a *late fusion* architecture that integrates projected visual features with an LLM only at the final stage of prediction. Given an image and caption during training, a pre-trained multimodal encoder maps the image to a joint image–text space,

* Equal contribution.

which is then projected to a sequence of pseudo-text embeddings $z_{1:n}^v$. In parallel, the input text is processed by a pre-trained LLM to produce token embeddings $z_{1:k}^x$. A late-stage attention-like mechanism allows $z_{1:k}^x$ to attend to $z_{1:n}^v$ once, immediately before prediction, rather than feeding visual tokens into the LLM stack. This design keeps the LLM focused on language while enabling access to visual information when helpful.

Second, at inference, we introduce *multi-image* evidence (see Figure 1). Since paired images are not available at test time, we generate k images from the input text using a *distilled* text-to-image generator with batched, parallel sampling to minimize overhead. Each generated image is processed through the late-fusion module to produce a probability distribution. We also compute a text-only distribution. We then aggregate these $k+1$ distributions using entropy-aware weighting, allowing confident predictions to dominate while preserving the text-only path when visual input is unhelpful.

We evaluate on object commonsense (Wang et al., 2023), visual commonsense (ImageNetVC Xia et al., 2023), and standard language benchmarks (Dubey et al., 2024; Touvron et al., 2023; Team et al., 2023; Almazrouei et al., 2023). Our method substantially outperforms LLMs and prior VaLMs on visual commonsense, matches VLMs on vision-heavy tasks, and, when applied to strong LLMs such as LLaMA 3, also improves text-only performance, while adding only modest test-time overhead due to lightweight, batched generation.

2 Related Work

Large Language and Vision Models. LLMs achieve strong performance on text-based reasoning but struggle with visual understanding (Team et al., 2024). Vision–Language Models (VLMs) (Liu et al., 2023a; Dai et al., 2023; Cha et al., 2024) address this gap and excel at multimodal tasks such as VQA, image captioning, and visual commonsense reasoning (Xia et al., 2023; Li et al., 2023b; Jin et al., 2024), yet often degrade on purely textual commonsense reasoning.

Visually-Augmented Language Models. VaLMs augment text-only LMs with visual inputs. Some retrieve related images and feed them to the LM (Tan and Bansal, 2020; Lu et al., 2022; Wang et al., 2023), while others distill visual knowledge from multimodal models such as CLIP (Radford

et al., 2021) or BLIP-2 (Li et al., 2023a) into LMs (Tang et al., 2021; Zhang et al., 2024; Guo et al., 2023; Li et al., 2023b; Cui et al., 2024). Diffusion-based approaches like Z-LaVi (Yang et al., 2022) generate visuals for possible label predictions, whereas our method generates images directly from the input text. LiVE (Tang et al., 2023) adds a vision–text fusion layer into the LM, and iNLG (Zhu et al., 2023) uses visual prefixes to guide generation. In contrast, we keep the LLM unchanged, apply late fusion between its output and an image encoding, and aggregate predictions from multiple generated images via simple averaging, enabling diverse visual experts to guide towards confident outputs.

3 Method

Our method, LAMI (**L**ate **M**ulti-**I**mage fusion), enhances LMs with visual cues to improve object and visual commonsense reasoning while preserving text-only performance. We train with (i) image–caption pairs and (ii) text with a synthetically generated image from a text-to-image model. At inference, we generate k images for a prompt, process each with the model, and aggregate their predictions using confidence-based weighting.

3.1 Visually Enhanced Language Model

The architecture, illustrated in Fig. 2, consists of four components: a frozen pre-trained LLM, a frozen pre-trained vision encoder, a trainable Visual Token Projector (VTP), and a trainable Late Fusion Attention Layer (LFAL).

Given an image v and caption $x = (x_{(1)}, \dots, x_{(n_x)})$, the training objective is:

$$\max_{\theta} \log P_{\theta}(x_{(t)} | x_{(<t)}, v). \quad (1)$$

The vision encoder extracts patch features $z^v \in \mathbb{R}^{n_v \times d_v}$, which the VTP maps to pseudo-text embeddings via

$$u^v = W_1 \sigma(W_2 z^v), \quad u^v \in \mathbb{R}^{n_v \times d_x}.$$

The LFAL fuses u^v with text embeddings $z_{(<t)}^x$ by setting $K = V = [u^v; z_{(<t)}^x]$ and $Q = z_{(<t)}^x$, enabling text tokens to attend once to visual tokens before projecting to vocabulary logits.

3.2 Visually Driven Inference

Since paired images are unavailable at inference, we generate k images $\{v_i\}$ from the prompt using

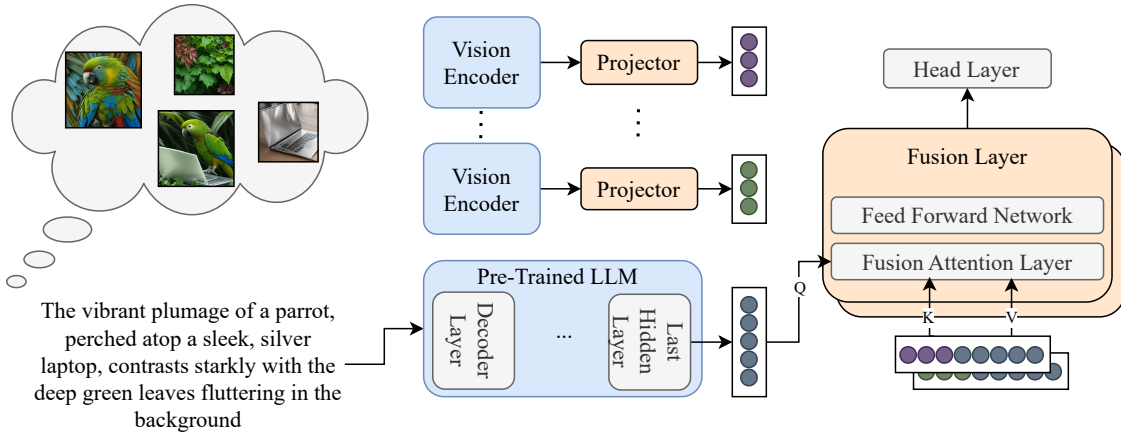


Figure 2: **Overview of LAMI**. Multiple images are generated from the input text and independently encoded by a frozen vision encoder, then projected to pseudo-text tokens. In parallel, the text is processed by a frozen pre-trained LLM. A trainable late-fusion attention layer allows the LLM’s final text representations to attend to the projected visual tokens, combining both modalities before the prediction head. Blue: frozen; orange: trainable.

Model	Mem. Color	Color Terms	Obj. Shape	Rel. Size
Masked Language Models				
BERT	31.6	30.7	28.1	38.1
BERT (FT)	33.9	31.5	21.5	35.7
Vokenization*	14.2	20.0	43.2	72.4
X-adapter*	64.1	60.0	-	-
LAMI ‡	74.5	72.5	67.3	78.4
Causal Language Models				
GPT-2	32.4	34.6	44.5	43.1
GPT-2 (FT)	33.3	34.9	39.3	38.2
LIVE‡	49.6	46.7	41.5	66.7
Z-LaVI†	50.4	49.2	64.4	76.8
VaLM* ($k=4$)	54.0	52.7	62.8	85.0
VaLM* ($k=8$)	58.6	50.2	59.4	62.4
LAMI ‡	72.5	69.2	66.8	85.5

Table 1: Object commonsense results. * retrieves images, † generates images during inference.

a distilled text-to-image generator with batched, parallel sampling. Each image yields a distribution $p_i = P_\theta(x_t | x_{<t}, v_i)$, and we also compute the text-only distribution $p_0 = P_\theta(x_t | x_{<t})$.

We ensemble the image-based predictions $p_{\text{img}} = \frac{1}{k} \sum_{i=1}^k p_i$, then weight them by a normalized CLIP score $f(\bar{x}_i, v_i)$:

$$p_{\text{final}} = \sum_{i=1}^k f(\bar{x}_i, v_i) p_i + (1 - f(\bar{x}_i, v_i)) p_0. \quad (2)$$

High-alignment images are trusted more, while low-alignment ones fall back to the text-only LLM.

4 Results

We first evaluate LAMI on object commonsense, comparing against prior visually-augmented LMs and ablating our contributions. We then scale to

larger models and evaluate on visual commonsense, reasoning, and reading comprehension. Next, we compare LAMI against a text-only baseline with matched inference-time budget. Finally, we analyze the effect of the number of generated images, compare integrating generated images against integrating multimodal embeddings, and ablate the CLIP-based fusion strategy. Additional analyses on generation vs. retrieval are provided in Sec. B.

Evaluation Benchmarks. We follow the zero-shot benchmark of Wang et al. (2023) for QA on object color, shape, and size, using Memory Color (Norlund et al., 2021) and Color Terms (Bruni et al., 2012) for color, Vi-ComTe (Zhang et al., 2022a) for shape, and the size dataset of Bagherinezhad et al. (2016), adhering to Wang et al. (2023) guidelines.

Comparison to Visually Augmented. Following the baselines, we start by studying weaker LMs, focusing on masked LMs (BERT) and causal LMs (GPT-2), both lacking visual commonsense (e.g., failing to answer *What is the color of a banana?*). For GPT-2 we use zero-shot accuracy; for BERT we follow Zhang et al. (2024), masking after the last word and predicting the masked token. To ensure a fair comparison, in this experiment LAMI is trained only on the Visual Genome (VG) dataset (Krishna et al., 2016), which contains an equal or smaller multimodal data than the baselines. Further implementation and baseline details are provided in the Appendix, Secs. A.1, A.2.

Table 1 shows that LAMI substantially improves performance across all tasks, with only a marginal gain over VaLM on Relative Size (85.0 vs. 85.5).

Method	Mem. Color	Color Terms	Obj. Shape	Rel. Size
GPT-2 (Base)	32.4	34.6	44.5	43.1
E-F.	49.1	45.3	40.3	70.1
E-F. + M	55.5	52.1	41.2	75.5
I-F.	62.8	59.3	60.0	77.2
I-F. + M	69.7	67.8	63.0	81.1
L-F.	65.1	62.2	63.5	80.2
L-F. + M (Ours)	72.5	69.2	66.8	85.5

Table 2: Ablation studies on fusion strategies and multi-image generation. Abbreviations: E-F.=Early Fusion, I-F.=Intermediate Fusion, L-F.=Late Fusion, M=Multi-Image Generation.

We attribute these improvements to integrating multiple generated images via late fusion, which is more robust than baselines that rely on a single image, early fusion, or simple probability summation.

To validate our assumptions about the contributing factors of our proposed components, we conduct ablation experiments. Table 2 reports results for seven configurations of the GPT-2 model: early fusion following iNLG (Zhu et al., 2023) and LiVE (Tang et al., 2023), early fusion with multi-image generation, intermediate fusion as in VaLM (Wang et al., 2023), intermediate fusion with multi-image generation, late fusion alone, and our complete approach combining late fusion with multi-image generation.

Results suggest that both architectural choices substantially improve performance, and that removing them reduces our method to levels comparable with the baselines. Multi-image generation consistently yields gains across all fusion strategies, particularly on color and relative size reasoning. Late fusion outperforms early and intermediate fusion, especially on shape-related tasks. The combination of late fusion and multi-image generation achieves the best overall performance.

Comparison to VLMs. Next, we evaluate LAMI on visual commonsense, commonsense reasoning, and reading comprehension using more advanced models, following the evaluation settings from Touvron et al. (2023). Detailed benchmarks and implementation settings are provided in Appendix, Secs. A.3 and A.1. Results provided in Tab. 4, LAMI consistently improves LMs of all sizes: small (GPT-2), medium (Gemma-2B), and large (Vicuna-7B-V1.5, Llama3-8B, Llama3-8B-Instruct). Unlike VLMs such as InstructBLIP and Llava-Next, which often improve visual commonsense at the cost of text-task performance, LAMI enhances visual commonsense while also maintaining or improving text-based results. This highlights

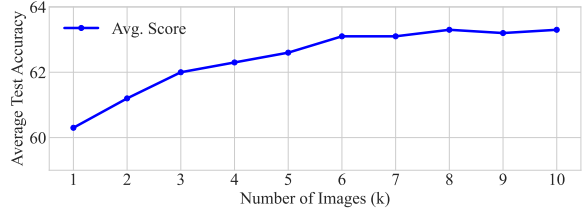


Figure 3: Average performance as a function of the number of generated images.

Method	VC	CR	RC
Gemma-2B	45.6	63.8	48.8
Gemma-2B (Best-of- N)	47.8	64.8	49.0
LAMI (Gemma-2B)	50.1	65.1	48.9

Table 3: Comparison of LAMI against Best-of- N sampling with matched inference runtime on Gemma-2B.

a key strength of late fusion: *it adds visual capability without sacrificing language reasoning.*

Inference-Time Compute Baseline. Since LAMI introduces additional inference cost through image generation, a natural question is whether comparable gains could be achieved by simply spending more compute on text-only decoding. To test this, we compare against a Best-of- N strategy: the model generates N independent completions and selects the one with the highest average token log-likelihood, with N calibrated so that total runtime matches LAMI ($k = 6$) on Gemma-2B. As shown in Tab. 3, Best-of- N yields modest gains on commonsense reasoning and reading comprehension but fails to close the visual commonsense gap, confirming that LAMI’s improvements stem from grounded visual evidence rather than additional compute alone.

Effect of k . In Figure 3, we analyze the effect of the number of generated images during inference, varying k from 1 to 10 on a validation set drawn from Color (Xia et al., 2023), PIQA (Bisk et al., 2019), and BoolQ (Clark et al., 2019). Performance improves as k increases, saturating around $k \approx 6$. Since image generation at inference is costly, we note that even $k = 3$ yields improvements.

Images vs. Embeddings. One might argue that multimodal representations could inject visual information instead of synthetic images (Guo et al., 2023). This raises the question: *do synthetically generated images hold more information than multimodal text representations?* To address this, we compare LAMI (Gemma-2B) against CLIP text

Model	Base	VC	CR	RC	Avg.
<i>Small-Scale Models</i>					
GPT-2	-	30.3	46.1	30.5	35.6
LAMI	GPT-2	38.6	46.7	32.2	39.2
<i>Mid-Scale Models</i>					
Gemma-2B	-	45.6	63.8	48.8	52.7
LAMI	Gemma-2B	50.1	65.1	48.9	54.7
<i>Large-Scale Models</i>					
Vicuna-7B	-	45.1	57.6	57.5	53.4
InstructBLIP*	Vicuna-7B	50.1	52.6	53.6	52.1
Llava-Next*	Vicuna-7B	50.3	54.5	54.7	53.1
LAMI	Vicuna-7B	48.6	58.8	57.9	55.1
Llama3-8B	-	52.0	72.0	57.9	60.6
LAMI	Llama3-8B	55.0	72.9	58.0	62.0
<i>Large-Scale Instruct Models</i>					
Llama3-8B-Instruct	-	53.0	71.6	59.2	61.2
Llava-Next*	Llama3-8B-Inst.	56.5	70.8	54.8	60.7
LAMI	Llama3-8B-Inst.	55.6	71.7	60.9	62.7
Qwen-2.5-7B-Instruct	-	54.8	75.6	63.0	64.4
Qwen2-VL-7B-Instruct*	Qwen-2.5-7B-Inst.	59.0	74.4	57.9	63.7
LAMI	Qwen-2.5-7B-Inst.	57.8	75.9	64.4	66.0

Table 4: Results on visual commonsense (VC), commonsense reasoning (CR), and reading comprehension (RC). * = VLM trained on large-scale image-text data.

Method	VC	CR	RC	Time (ms)
Gemma-2B	45.6	63.8	48.8	20
CLIP Embedding	<u>47.9</u>	<u>64.7</u>	<u>48.9</u>	<u>58</u>
Generated Images	50.1	65.1	48.9	389

Table 5: Performance comparison of image generation, CLIP text embedding, and baseline (Gemma-2B).

embeddings, maintaining identical architecture, datasets, and implementation while replacing only the visual representations. Following Guo et al. (2023), we extract noun entities via part-of-speech tagging before CLIP encoding, as full-text embeddings yielded poor performance. We evaluate under the settings in A.3, reporting results in Tab. 5. While CLIP embeddings provide a computationally cheaper alternative that still improves over baseline, generated images outperform both CLIP embeddings and baseline across all tasks.

CLIP-Fusion Ablation. Lastly, we ablate the CLIP-based fusion strategy described in Sec. 3.2 by comparing it against purely confidence-based aggregation methods. Specifically, we evaluate: (i) Single Image, using a single generated image without aggregation; (ii) Average Logits, averaging the prediction distributions across all k generated images; and (iii) Max Confidence, selecting the prediction with the lowest entropy among the k image-conditioned distributions. All variants use Gemma-2B with $k = 6$ and are evaluated on Color, PIQA, and BoolQ. Results are reported in Tab. 6. CLIP-fusion consistently outperforms all

Method	Color	PIQA	BoolQ
Single Image	40.8	76.1	66.1
Average Logits	44.6	76.5	66.5
Max Confidence	43.0	76.1	66.9
LAMI (CLIP-fusion)	45.4	77.7	67.0

Table 6: Ablation of aggregation strategies. CLIP-fusion outperforms purely confidence-based alternatives.

confidence-based alternatives.

5 Qualitative Analysis and Failure Cases

We provide representative examples using Llama-3 to illustrate the behavior of LAMI .

Success case. *How many humps does a Bactrian camel have?* Llama-3 predicts *one*, confusing Dromedary and Bactrian species. LAMI generates images of two-humped camels, correcting the prediction to *two*.

Robustness to negation. *Which color is not on a stop sign?* The generator produces a red stop sign, which is misleading under negation. However, the low CLIP alignment score suppresses the visual path, and LAMI falls back to the text-only prediction, correctly outputting *blue*.

Failure case. *What material holds the Sword of Damocles?* Llama-3 predicts *a thin rope*. LAMI generates images depicting a metal chain, a visually plausible but incorrect depiction (the correct answer is *a single horse hair*). The high alignment score causes the visual path to override the text prior. Such failures are more likely for abstract or legendary concepts where text-to-image generators lack faithful grounding.

6 Conclusions

We introduce LAMI , a framework for enhancing the visual commonsense capabilities of LLMs without compromising their text reasoning. Unlike prior work, we focus on two overlooked but crucial aspects of visual augmentation: (i) leveraging multiple generated images to capture diverse visual evidence, and (ii) applying late fusion over text and visual features for robust integration. Comprehensive evaluation shows that both components are essential, with LAMI achieving strong gains on visual commonsense tasks while preserving text-based performance.

7 Limitations

Visually augmented techniques incur additional cost, as image generation is slower than text decoding. For instance, Gemma-2B requires ~ 20 ms per token, while our approach adds ~ 50 ms per image. Efficiency can be improved through parallelized image generation, which bounds latency, or by applying visual conditioning as initial context or selectively during decoding.

Nevertheless, this trade-off reflects a broader trend: scaling test-time compute enhances output quality (Wei et al., 2023; Yao et al., 2023; Snell et al., 2024), consistent with the “no-free-lunch” principle that stronger performance demands longer runtime. In this light, visually augmented reasoning represents a principled form of test-time scaling, likely to become more natural within agentic frameworks. Thus, we view it not as a stopgap but as a practical and forward-looking direction for visual reasoning.

Finally, we have not yet evaluated LAMI on the latest reasoning LLMs (DeepSeek-AI et al., 2025) or scaled training beyond 8B parameters due to computational constraints. We believe that reasoning with augmented images has the potential to advance state-of-the-art performance further.

8 Ethical Considerations

Our method inherits risks from its component models: the base LLM, vision encoder (CLIP), and text-to-image generator (SDXL-turbo). Key considerations include:

Hallucinations and Factual Grounding. Both the LLM and text-to-image generator may produce outputs not grounded in facts. Generated images may depict incorrect visual information, potentially compounding reasoning errors.

Bias Propagation. Biases present in the pre-trained LLM, CLIP encoder, and image generator can propagate through our pipeline, leading to unfair or stereotypical visual representations that influence final predictions.

Computational Cost. While our model primarily uses pre-trained foundation models as part of our model design and only adapts a lightweight vision projector and a fusion layer, training such pre-trained models requires significant energy consumption. Further, inference time queries, which are performed many times, may be costly.

References

- Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, and 8 others. 2022. [Flamingo: a visual language model for few-shot learning](#). *Preprint*, arXiv:2204.14198.
- Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, M erouane Debbah,  tienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, and 1 others. 2023. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*.
- Hessam Bagherinezhad, Hannaneh Hajishirzi, Yejin Choi, and Ali Farhadi. 2016. [Are elephants bigger than butterflies? reasoning about sizes of objects](#). *Preprint*, arXiv:1602.00753.
- Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2019. [Piqa: Reasoning about physical commonsense in natural language](#). *Preprint*, arXiv:1911.11641.
- Elia Bruni, Gemma Boleda, Marco Baroni, and Nam Khanh Tran. 2012. [Distributional semantics in technicolor](#). In *Annual Meeting of the Association for Computational Linguistics*.
- Junbum Cha, Wooyoung Kang, Jonghwan Mun, and Byungseok Roh. 2024. [Honeybee: Locality-enhanced projector for multimodal llm](#). *Preprint*, arXiv:2312.06742.
- Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. [Quac: Question answering in context](#). *arXiv preprint arXiv:1808.07036*.
- Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. [Boolq: Exploring the surprising difficulty of natural yes/no questions](#). *arXiv preprint arXiv:1905.10044*.
- Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. [Think you have solved question answering? try arc, the ai2 reasoning challenge](#). *arXiv preprint arXiv:1803.05457*.
- Wanqing Cui, Keping Bi, Jiafeng Guo, and Xueqi Cheng. 2024. [More: Multi-modal retrieval augmented generative commonsense reasoning](#). *Preprint*, arXiv:2402.13625.
- Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. [Instructblip: Towards general-purpose vision-language models with instruction tuning](#). *Preprint*, arXiv:2305.06500.

- DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). *Preprint*, arXiv:2501.12948.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [Bert: Pre-training of deep bidirectional transformers for language understanding](#). *Preprint*, arXiv:1810.04805.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, and 1 others. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.
- Hangyu Guo, Kun Zhou, Wayne Xin Zhao, Qinyu Zhang, and Ji-Rong Wen. 2023. [Visually-augmented pretrained language models for nlp tasks without images](#). *Preprint*, arXiv:2212.07937.
- Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, Qiang Liu, Kriti Aggarwal, Zewen Chi, Johan Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and Furu Wei. 2023. [Language is not all you need: Aligning perception with language models](#). *Preprint*, arXiv:2302.14045.
- Woojeong Jin, Tejas Srinivasan, Jesse Thomason, and Xiang Ren. 2024. [Winoviz: Probing visual properties of objects under different states](#). *Preprint*, arXiv:2402.13584.
- Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. 2016. [Visual genome: Connecting language and vision using crowdsourced dense image annotations](#). *Preprint*, arXiv:1602.07332.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. [Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension](#). *Preprint*, arXiv:1910.13461.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. [Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models](#). *Preprint*, arXiv:2301.12597.
- Lei Li, Jingjing Xu, Qingxiu Dong, Ce Zheng, Qi Liu, Lingpeng Kong, and Xu Sun. 2023b. Can language models understand physical concepts? *arXiv preprint arXiv:2305.14057*.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. [Microsoft coco: Common objects in context](#). *Preprint*, arXiv:1405.0312.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. In *NeurIPS*.
- Yujie Lu, Wanrong Zhu, Xin Eric Wang, Miguel Eckstein, and William Yang Wang. 2022. [Imagination-augmented natural language understanding](#). *Preprint*, arXiv:2204.08535.
- Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. [Pointer sentinel mixture models](#). *Preprint*, arXiv:1609.07843.
- Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*.
- Tobias Norlund, Lovisa Hagström, and Richard Johansson. 2021. [Transferring knowledge from vision to language: How to achieve it and how to measure it?](#) *Preprint*, arXiv:2109.11321.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.
- Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2016. [Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models](#). *Preprint*, arXiv:1505.04870.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. [Learning transferable visual models from natural language supervision](#). *Preprint*, arXiv:2103.00020.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. [Language models are unsupervised multitask learners](#).
- Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*.

- Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.
- Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*.
- Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. 2023. [Adversarial diffusion distillation](#). *Preprint*, arXiv:2311.17042.
- Christoph Schuhmann and Peter Bevan. 2023. [220k-gpt4vision-captions-from-lvis](https://huggingface.co/datasets/laion/220k-GPT4Vision-captions-from-LIVIS). <https://huggingface.co/datasets/laion/220k-GPT4Vision-captions-from-LIVIS>.
- Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. [Conceptual captions: A cleaned, hypernamed, image alt-text dataset for automatic image captioning](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.
- Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Unsupervised commonsense question answering with self-talk. *arXiv preprint arXiv:2004.05483*.
- Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. [Scaling llm test-time compute optimally can be more effective than scaling model parameters](#). *Preprint*, arXiv:2408.03314.
- Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*.
- Hao Tan and Mohit Bansal. 2020. [Vokenization: Improving language understanding with contextualized, visual-grounded supervision](#). *Preprint*, arXiv:2010.06775.
- Tianyi Tang, Yushuo Chen, Yifan Du, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. 2023. [Learning to imagine: Visually-augmented natural language generation](#). *Preprint*, arXiv:2305.16944.
- Zineng Tang, Jaemin Cho, Hao Tan, and Mohit Bansal. 2021. Vidlankd: Improving language understanding via video-distilled knowledge transfer. *Advances in Neural Information Processing Systems*, 34:24468–24481.
- Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, and 89 others. 2024. [Gemma: Open models based on gemini research and technology](#). *Preprint*, arXiv:2403.08295.
- MosaicML NLP Team and 1 others. 2023. [Introducing mpt-7b: A new standard for open-source, commercially usable llms, 2023](#). *URL www.mosaicml.com/blog/mpt-7b*. Accessed, pages 05–05.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. [Llama 2: Open foundation and fine-tuned chat models](#). *Preprint*, arXiv:2307.09288.
- Weizhi Wang, Li Dong, Hao Cheng, Haoyu Song, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. 2023. [Visually-augmented language modeling](#). *Preprint*, arXiv:2205.10178.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. [Chain-of-thought prompting elicits reasoning in large language models](#). *Preprint*, arXiv:2201.11903.
- Heming Xia, Qingxiu Dong, Lei Li, Jingjing Xu, Tianyu Liu, Ziwei Qin, and Zhifang Sui. 2023. [Imagenetvc: Zero- and few-shot visual commonsense evaluation on 1000 imagenet categories](#). *Preprint*, arXiv:2305.15028.
- Yue Yang, Wenlin Yao, Hongming Zhang, Xiaoyang Wang, Dong Yu, and Jianshu Chen. 2022. [Z-lavi: Zero-shot language solver fueled by visual imagination](#). *Preprint*, arXiv:2210.12261.
- Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. [Tree of thoughts: Deliberate problem solving with large language models](#). *Preprint*, arXiv:2305.10601.
- Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*.
- Chenyu Zhang, Benjamin Van Durme, Zhuowan Li, and Elias Stengel-Eskin. 2022a. [Visual commonsense in pretrained unimodal and multimodal models](#). *Preprint*, arXiv:2205.01850.
- Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022b. [Opt: Open pre-trained transformer language models](#). *Preprint*, arXiv:2205.01068.

Xinyun Zhang, Haochen Tan, Han Wu, and Bei Yu. 2024. [Towards versatile and efficient visual knowledge integration into pre-trained language models with cross-modal adapters](#). *Preprint*, arXiv:2305.07358.

Wanrong Zhu, An Yan, Yujie Lu, Wenda Xu, Xin Eric Wang, Miguel Eckstein, and William Yang Wang. 2023. [Visualize before you write: Imagination-guided open-ended text generation](#). *Preprint*, arXiv:2210.03765.

A Experimental Settings

A.1 Implementation Details

We use CLIP-ViT/32 (Radford et al., 2021) as the vision encoder and for computing text-image alignment scores, with SDXL-turbo (Sauer et al., 2023) for image generation. Training employed a two-stage pipeline on four A100 GPUs: 40K iterations (batch size 256, lr 5×10^{-4}) followed by 10K fine-tuning iterations (batch size 128, lr 5×10^{-5}) using AdamW. Only LFAL and VTP were trained while other components remained frozen. Training required 192 GPU hours for Llama-3, 90 for Gemma-2B/OPT-2.7B, and 50 for GPT-2.

We trained on Visual Genome Regions (5.4M images) (Krishna et al., 2016), Laion-220K (Schuhmann and Bevan, 2023), and 2% of Wikitext-103 (Merity et al., 2016). For Wikipedia texts, we synthetically generated corresponding images to simulate inference conditions. During inference, we generate 6 images per sample unless specified otherwise.

A.2 VaLM Baseline Details

We compared our method with several VaLMs designed to improve visual commonsense: Vokenization (Tan and Bansal, 2020) and X-adapter (Zhang et al., 2024) (BERT-based); Z-LaVI (Yang et al., 2022) (GPT-neo-1.3B (Gao et al., 2020)); LIVE (Tang et al., 2023) (BART-based (Lewis et al., 2019)); and VaLM (Wang et al., 2023) (GPT-2). We also compared against pure LMs (BERT and GPT-2) and their fine-tuned versions trained on the same data without images.

For fair comparison, in this comparison, we trained our method only on the VG dataset. All baselines except VaLM either trained on VG during pretraining or retrieved VG images during inference. Specifically, Vokenization and X-adapter use COCO (Lin et al., 2015) and VG; LIVE incorporates COCO, VG, CC3M (Sharma et al., 2018), and Flickr30k (Plummer et al., 2016). Z-LaVI,

a zero-shot model, was employed with VG and Bing Image Search collections. VaLM trains GPT-2 from scratch and lacks publicly available weights, so we report their published results.

For the binary Relative Size test, GPT-2, BERT, Vokenization, and LIVE exhibited strong yes/no bias. We addressed this by fine-tuning open-weight models on 3,200 yes/no questions from ViComTe size (Zhang et al., 2022a) for three epochs with learning rate 5×10^{-5} .

A.3 Visual Commonsense, Commonsense Reasoning, and Reading Comprehension Benchmarks

We evaluate LAMI across three benchmark categories following Touvron et al. (2023). For visual commonsense, we use ImageNetVC (Xia et al., 2023). For Commonsense Reasoning, we evaluate on PIQA (Bisk et al., 2019), SIQA (Sap et al., 2019), HellaSwag (Zellers et al., 2019), Winogrande (Sakaguchi et al., 2021), ARC easy and challenge (Clark et al., 2018), OpenBookQA (Mihaylov et al., 2018), and CommonsenseQA (Talmor et al., 2018). For Reading Comprehension, we use BoolQ (Clark et al., 2019), SQuAD 2.0 (Rajpurkar et al., 2018), and QuAC (Choi et al., 2018). For base models, we measure accuracy by selecting the answer with highest likelihood from candidate sets following Schwartz et al. (2020). For instruct models (Vicuna-7B-V1.5, Llama3-8B-Instruct), we use instruction-style prompts with top-1 accuracy following Xia et al. (2023). SQuAD and QuAC use exact match scoring per Ouyang et al. (2022), while BoolQ employs zero-shot binary selection between yes/no tokens.

B Generation vs. Retrieval

Prior VaLM baselines such as VaLM (Wang et al., 2023) and X-adapter (Zhang et al., 2024) rely on image retrieval. While our main results (Tab. 1) already show that LAMI outperforms these methods, we further isolate the effect by replacing our generation module with VaLM’s retrieval mechanism, keeping all other components identical. On GPT-2, this substitution reduces Memory Color from 72.5 to 65.5 and Color Terms from 69.2 to 62.8, confirming that generation offers superior input specificity and diversity.