Multimodal Chain of Continuous Thought for Latent-Space Reasoning in Vision-Language Models

Tan-Hanh Pham[1,2,†], **Chris Ngo**[3]

[1]Harvard Medical School, Harvard University
[2]Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital
[3]Knovel Engineering Lab, Singapore
[†]Corresponding author: tpham33@mgh.harvard.edu

Abstract

Many reasoning techniques for large multimodal models adapt language model approaches, such as Chain-of-Thought (CoT) prompting, which express reasoning as word sequences. While effective for text, these methods are suboptimal for multimodal contexts, struggling to align audio, visual, and textual information dynamically. In this work, we propose the Multimodal Chain of Continuous Thought (MCOUT), which enables reasoning directly in a joint latent space rather than in natural language. In MCOUT, the reasoning state is represented as a continuous hidden vector, iteratively refined and aligned with visual and textual embeddings, inspired by human reflective cognition. We develop two variants: MCOUT-Base, which reuses the language model's last hidden state as the continuous thought for iterative reasoning, and MCOUT-Multi, which integrates multimodal latent attention to strengthen cross-modal alignment between visual and textual features. Experiments on benchmarks including MMMU, ScienceQA, and MMStar show that MCOUT consistently improves multimodal reasoning, yielding up to 8.23% accuracy gains over strong baselines and improving BLEU scores up to 8.27% across multiple-choice and open-ended tasks. These findings highlight latent continuous reasoning as a promising direction for advancing LMMs beyond language-bound CoT, offering a scalable framework for human-like reflective multimodal inference.

1 Introduction

The development of reasoning in VLMs is crucial for tasks like VQA and multimodal reasoning, with methods ranging from prompting to reinforcement learning. Chain-of-Thought (CoT) prompting [1] and its variants, including Tree of Thoughts (ToT) [2], and Graph of Thoughts (GoT) [3], have advanced reasoning in LLMs by structuring intermediate steps, but remain computationally heavy and less effective in multimodal contexts due to alignment challenges and token overhead. Training-based techniques, including RLHF [4], GRPO [5], and reasoning functions like RARL [6], improve logical consistency and multimodal alignment, yet still struggle with static embeddings and inefficient integration of visual and textual data. More recently, latent reasoning paradigms shift reasoning from token sequences to continuous latent spaces, enabling efficiency and backtracking while reducing computational cost. Approaches such as COCONUT [7], LaRS [8], and hierarchical latent refinement [9] have proven effective for LLMs, though their application to VLMs remains limited.

Emerging VLM research has begun adapting latent reasoning, with frameworks such as multimodal CoT with diffusion [10], MMaDA [11], Mirage [12], stacked latent attention [13], multimodal latent language modeling [14], Corvid [15], and GCoT [16] improving alignment, dynamic reasoning, and reducing hallucinations. However, many remain constrained by token-based reasoning or static vision

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Efficient Reasoning.

features. To address these gaps, we propose MCOUT, a latent reasoning framework tailored for VLMs. MCOUT-Base uses the language model's last hidden state as a continuous thought, while MCOUT-Multi integrates it with image embeddings via multimodal latent attention for dynamic alignment. By iteratively refining thoughts in a continuous latent space, MCOUT mimics human reflective reasoning, offering efficiency, scalability, and robustness in vision-language reasoning.

2 Methodology

2.1 Model Architecture

The MCOUT framework is built upon a vision-language model, SilVar [17], comprising a pre-trained visual encoder $\mathcal{V}$ and a language model $\mathcal{L}$ (Appendix A). We use CLIP [18] as the visual encoder $\mathcal{V}$, which processes input images $\mathbf{x}_v \in \mathbb{R}^{H \times W \times C}$ to produce visual embeddings $\mathbf{e}_v \in \mathbb{R}^{S_v \times D}$, where S_v is the sequence length of visual tokens and D is the embedding dimension. For the language model $\mathcal{L}$, we employ Llama 3.2 1B, which processes tokenized text inputs $\mathbf{x}_t$ to generate contextual embeddings $\mathbf{e}_t \in \mathbb{R}^{S_t \times D}$, where S_t is the sequence length of text tokens.

For MCOUT-Multi, the core component is the multimodal latent attention module, which integrates the language model's last hidden state $\mathbf{h}_l \in \mathbb{R}^{B \times D}$ for a batch of B samples with multimodal input embeddings $\mathbf{e}_m \in \mathbb{R}^{B \times S_m \times D}$ (for images, $\mathbf{e}_m = \mathbf{e}_v$). The module projects $\mathbf{h}_l$ into a query space, applies multi-head attention with $N_h = 8$ heads to attend to $\mathbf{e}_m$, and normalizes the output to produce a thought embedding:

$$\mathbf{h}_t = \text{Norm}(\text{Proj}_{\text{back}}(\text{MultiHeadAttn}(\text{Proj}(\mathbf{h}_l), \mathbf{e}_m^\top))) \in \mathbb{R}^{B \times 1 \times D}, \tag{1}$$

where $\text{Proj} : \mathbb{R}^D \to \mathbb{R}^D$ and $\text{Proj}_{\text{back}} : \mathbb{R}^D \to \mathbb{R}^D$ are linear projections, and Norm denotes layer normalization. This process enriches $\mathbf{h}_t$ with visual context for cross-modal alignment. In contrast, MCOUT-Base bypasses this module, directly using the last hidden state as the thought embedding:

$$\mathbf{h}_t = \mathbf{h}_l \in \mathbb{R}^{B \times 1 \times D}. \tag{2}$$

MCOUT-Base relies on the language model's internal state for reasoning, while MCOUT-Multi enhances it through multimodal fusion, mimicking human reflective reasoning by validating thoughts against input embeddings.

2.2 Multimodal Latent Reasoning

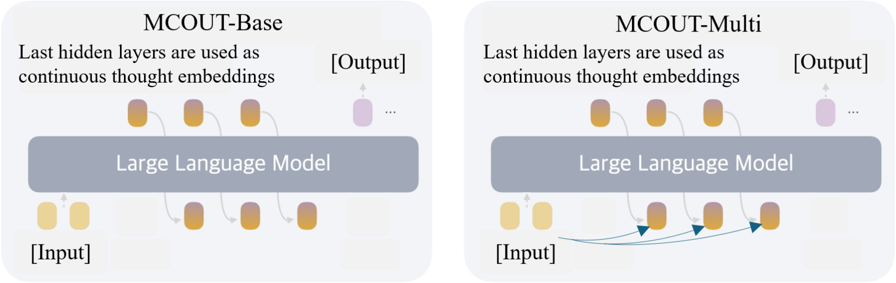
MCOUT-Base
Last hidden layers are used as continuous thought embeddings
[Output]
...
Large Language Model
[Input]
MCOUT-Multi
Last hidden layers are used as continuous thought embeddings
[Output]
...
Large Language Model
[Input]

Figure 1: Comparison between two Chain of Continuous Thought approaches: MCOUT-Base (left) vs. MCOUT-Multi (right).

The MCOUT framework performs reasoning by iteratively generating continuous thought representations in a latent space, inspired by human cognition, where intermediate thoughts are validated against input data for coherence, as shown in Figure 1. Given preprocessed interleaved input embeddings $\mathbf{e}_{\text{inter}} \in \mathbb{R}^{B \times S_{\max} \times D}$ and an attention mask $\mathbf{m} \in \{0, 1\}^{B \times S_{\max}}$ for a batch of B samples with maximum sequence length $S_{\max}$, the language model $\mathcal{L}$ computes hidden states:

$$\mathbf{h} = \mathcal{L}(\mathbf{e}_{\text{inter}}, \mathbf{m}) \in \mathbb{R}^{B \times S_{\max} \times D}. \tag{3}$$

The last hidden state for each sample is extracted by selecting the hidden state corresponding to the last non-padded token:

$$\mathbf{h}_l = \mathbf{h}[\cdot, \text{argmax}(\mathbf{m}, \text{dim} = 1) - 1, \cdot] \in \mathbb{R}^{B \times D}. \tag{4}$$

For N_t latent reasoning steps, MCOUT iteratively produces thought embeddings $\mathbf{h}_t^{(k)}$ for $k = 1, \ldots, N_t$. As mentioned, we explore two approaches: MCOUT-Base directly feeds the last hidden state to the language model N_t times, while MCOUT-Multi combines the last hidden state with input embeddings before feeding the resulting thought embedding to the language model:

- In MCOUT-Base:
$$\mathbf{h}_t^{(k)} = \mathbf{h}_l^{(k-1)} \in \mathbb{R}^{B \times 1 \times D}, \tag{5}$$
- In MCOUT-Multi:
$$\mathbf{h}_t^{(k)} = \text{MultimodalLatentAttention}(\mathbf{h}_l^{(k-1)}, \mathbf{e}_m) \in \mathbb{R}^{B \times 1 \times D}. \tag{6}$$

Each thought embedding is appended to the input sequence:

$$\mathbf{e}_{\text{inter}}^{(k)} = [\mathbf{e}_{\text{inter}}^{(k-1)}, \mathbf{h}_t^{(k)}] \in \mathbb{R}^{B \times (S_{\max}+k) \times D}, \tag{7}$$

$$\mathbf{m}^{(k)} = [\mathbf{m}^{(k-1)}, \mathbf{1}_{B \times 1}] \in \{0, 1\}^{B \times (S_{\max}+k)}. \tag{8}$$

The updated sequence is fed back into the language model to compute the next hidden state, repeating for N_t iterations. In the final step ($k = N_t + 1$), the language model generates the output sequence ($\mathbf{x}_a$) using a standard generation process. The loss function for training combines an auxiliary loss for intermediate thoughts (weighted by μ) and the final output loss:

$$\mathcal{L}_{\text{total}} = \sum_{k=1}^{N_t} \mu \cdot \mathcal{L}_{\text{aux}}^{(k)} + \mathcal{L}_{\text{final}}, \tag{9}$$

where $\mathcal{L}_{\text{aux}}^{(k)}$ is the language modeling loss for the k-th thought, and $\mathcal{L}_{\text{final}}$ is the loss for the final output, computed using cross-entropy over the target tokens.

3 Experiment and Result

Table 1: Performance comparison of MCOUT and baseline models across different datasets.

Dataset	Models	Accuracy (%)	BLEU	Gain
ScienceQA	Baseline	56.17	51.48	–
	MCOUT-Base (N_t=5)	58.60	52.44	+4.33%, +1.87%
	MCOUT-Multi (N_t=5)	58.45	**52.60**	+4.05%, +2.18%
	MCOUT-Base (N_t=10)	**58.86**	52.31	+4.79%, +1.61%
	MCOUT-Multi (N_t=10)	58.20	52.27	+3.61%, +1.53%
MMMU	Baseline	25.44	25.44	–
	MCOUT-Base (N_t=5)	**27.53**	**27.54**	**+8.21**%, **+8.31**%
	MCOUT-Multi (N_t=5)	27.18	27.19	+6.79%, +6.82%
	MCOUT-Base (N_t=10)	27.52	27.54	+8.18%, +8.31%
	MCOUT-Multi (N_t=10)	27.36	27.37	+7.54%, +7.58%
MMStar	Baseline	25.13	25.14	–
	MCOUT-Base (N_t=10)	**26.13**	**26.14**	+3.98%, +3.98%
	MCOUT-Multi (N_t=10)	26.07	26.08	+3.74%, +3.74%
ScienceQA	Kosmos2 [19]	32.70	–	–
	InstructBLIP-7B [20]	54.10	–	–
MMMU	Kosmos2 [19]	23.70	–	–
	Qwen-VL [21]	29.60	–	–
MMStar	Kosmos2 [19]	24.90	–	–
	LLaVA-7B [22]	27.10	–	–
	OpenFlamingo-9B [23]	26.90	–	–

To evaluate the effectiveness of our MCOUT framework, we conducted experiments on four diverse vision-language datasets: VQAv2 [24], MMMU [25], ScienceQA [26], and MMStar [27]. Specifically, we used VQAv2 for pretraining, MMMU and ScienceQA for finetuning and evaluation, and MMStar solely for testing. The results in Table 1 show that both variants consistently outperform the baseline, with MCOUT-Base excelling on ScienceQA (58.86% accuracy, +4.79%) and MMMU (+8.21 accuracy, +8.31% BLEU), while MCOUT-Multi achieves the best BLEU score on ScienceQA (52.60, +2.18%) and comparable gains on MMStar (+3.74%). Despite having only 1B parameters, MCOUT surpasses larger models such as Kosmos-2 (1.7B) and InstructBLIP-7B, and approaches the performance of LLaVA-7B and OpenFlamingo-9B. (additional comparisons are provided in Appendix B). These results highlight MCOUT's efficiency in leveraging latent reasoning for diverse multimodal tasks, ranging from image-heavy science questions to fine-grained visual reasoning.

4 Multimodal Latent Reasoning Analysis

To compare MCOUT-Base and MCOUT-Multi, we analyzed their latent distributions (Figure 2). Before training, a large norm gap was observed on ScienceQA (**103.90** for the last hidden state vs. **26.48** for multimodal attention embeddings), risking unstable fusion in MCOUT-Multi. To address this, we introduced normalization layers to align the scales and stabilize the embeddings. As a result, for MCOUT-Base, which directly reuses the last hidden ($\mathbf{h}_t = \mathbf{h}_l$), the mean remains near zero (–0.02197) with stable variance ($\sigma \approx 2.23$), supporting consistent reasoning and accuracy gains (+4.79% on ScienceQA, +8.21% on MMMU). In contrast, MCOUT-Multi shows similar stability in the last hidden states but its mixed embeddings exhibit near-constant mean (0.0024) and low variance ($\sigma \approx 0.20$), indicating modality collapse. This static contribution limits cross-modal benefits, aligning its accuracy (58.45%) with MCOUT-Base (58.60%). These findings resonate with recent work on visual attention collapse [28–30], suggesting that low-variance multimodal embeddings diminish fusion effectiveness despite norm adjustments. In addition, we conducted an ablation study on the auxiliary weight μ (ranging from 0 to 1) in the MCOUT loss (Equation 9) with $N_t = 5$, and found that $\mu = 0.3$ achieved the best performance, improving ScienceQA accuracy by 4.33% and MMMU accuracy by 8.23%. Although auxiliary loss boosts multimodal reasoning, higher μ values diminish gains, $\mu = 0$ yields only moderate improvements, and training time increases as a trade-off.

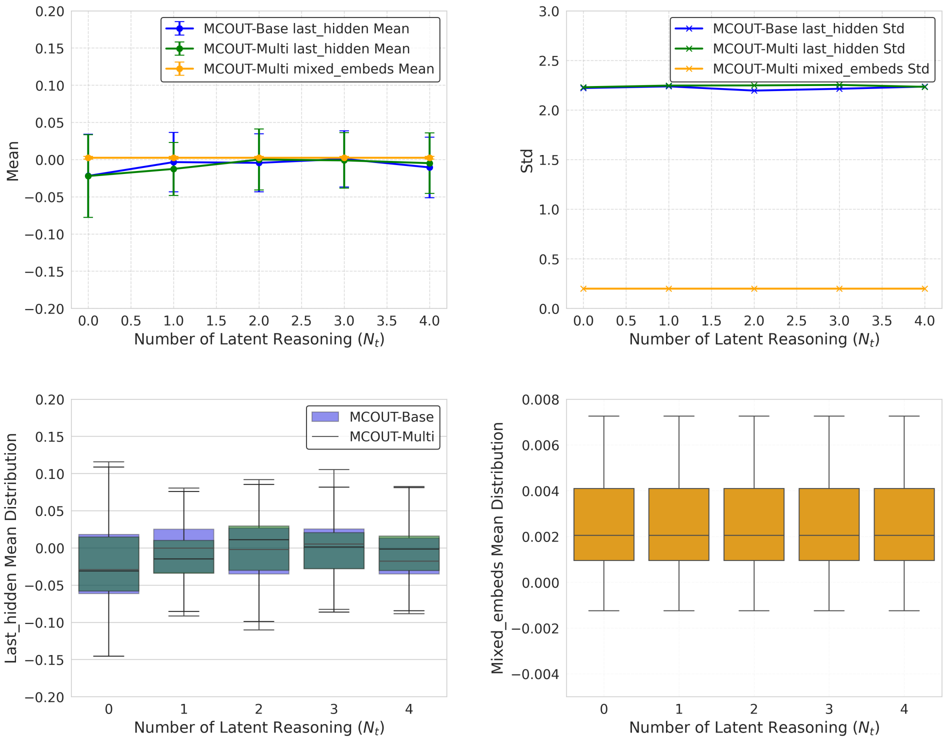

Figure 2: Latent distribution analysis of MCOUT-Base and MCOUT-Multi, showing mean and standard deviation of last hidden states and mixed embeddings across 100 samples with N_t=5.

5 Conclusion

In this work, we investigated multimodal reasoning for a small VLM through two key contributions: (1) building a 1B-parameter vision-language model, and (2) proposing the Multimodal Chain of Continuous Thought (MCOUT) framework, which employs a step-by-step reasoning process inspired by human reflection. MCOUT improves performance, achieving gains of up to 8.23% in accuracy on MMMU and 4.79% on ScienceQA. As a pioneering effort to explore multimodal continuous latent reasoning, our study provides a promising foundation for efficient multimodal reasoning. Despite these advances, aligning input embeddings with the final hidden layers remains a challenge, as it complicates multimodal alignment in MCOUT and increases training time.

References

[1] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[2] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023.

[3] Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, et al. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 17682–17690, 2024.

[4] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[5] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[6] Tan-Hanh Pham and Chris Ngo. Rarl: Improving medical vlm reasoning and generalization with reinforcement learning and lora under data and hardware constraints. *arXiv preprint arXiv:2506.06600*, 2025.

[7] Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*, 2024.

[8] Zifan Xu, Haozhu Wang, Dmitriy Bespalov, Xuan Wang, Peter Stone, and Yanjun Qi. Latent skill discovery for chain-of-thought reasoning. *arXiv preprint arXiv:2312.04684*, 2023.

[9] Guan Wang, Jin Li, Yuhao Sun, Xing Chen, Changling Liu, Yue Wu, Meng Lu, Sen Song, and Yasin Abbasi Yadkori. Hierarchical reasoning model. *arXiv preprint arXiv:2506.21734*, 2025.

[10] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023.

[11] Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. Mmada: Multimodal large diffusion language models. *arXiv preprint arXiv:2505.15809*, 2025.

[12] Zeyuan Yang, Xueyang Yu, Delin Chen, Maohao Shen, and Chuang Gan. Machine mental imagery: Empower multimodal reasoning with latent visual tokens. *arXiv preprint arXiv:2506.17218*, 2025.

[13] Haoqi Fan and Jiatong Zhou. Stacked latent attention for multimodal reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1072–1080, 2018.

[14] Yutao Sun, Hangbo Bao, Wenhui Wang, Zhiliang Peng, Li Dong, Shaohan Huang, Jianyong Wang, and Furu Wei. Multimodal latent language modeling with next-token diffusion. *arXiv preprint arXiv:2412.08635*, 2024.

[15] Jingjing Jiang, Chao Ma, Xurui Song, Hanwang Zhang, and Jun Luo. Corvid: Improving multimodal large language models towards chain-of-thought reasoning. *arXiv preprint arXiv:2507.07424*, 2025.

[16] Qiong Wu, Xiangcong Yang, Yiyi Zhou, Chenxin Fang, Baiyang Song, Xiaoshuai Sun, and Rongrong Ji. Grounded chain-of-thought for multimodal large language models. *arXiv preprint arXiv:2503.12799*, 2025.

[17] Tan-Hanh Pham, Trong-Duong Bui, Minh Luu Quang, Tan Huong Pham, Chris Ngo, and Truong Son Hy. Silvar-med: A speech-driven visual language model for explainable abnormality detection in medical imaging. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2984–2994, 2025.

[18] 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.

[19] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023.

[20] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023.

[21] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

[22] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

[23] Anas Awadalla, Irena Gao, Joshua Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo, March 2023.

[24] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[25] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*, 2024.

[26] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[27] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024.

[28] Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. See what you are told: Visual attention sink in large multimodal models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[29] Nicola Cancedda. Spectral filters, dark signals, and attention sinks. *arXiv preprint arXiv:2402.09221*, 2024.

[30] Mong Yuan Sim, Wei Emma Zhang, Xiang Dai, and Biaoyan Fang. Can vlms actually see and read? a survey on modality collapse in vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24452–24470, 2025.

[31] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

[32] Xiaocui Yang, Wenfang Wu, Shi Feng, Ming Wang, Daling Wang, Yang Li, Qi Sun, Yifei Zhang, Xiaoming Fu, and Soujanya Poria. Mm-bigbench: Evaluating multimodal models on multimodal content comprehension tasks. *arXiv preprint arXiv:2310.09036*, 2023.

[33] Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. Pandagpt: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355*, 2023.

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

Appendix

A Datasets and Training

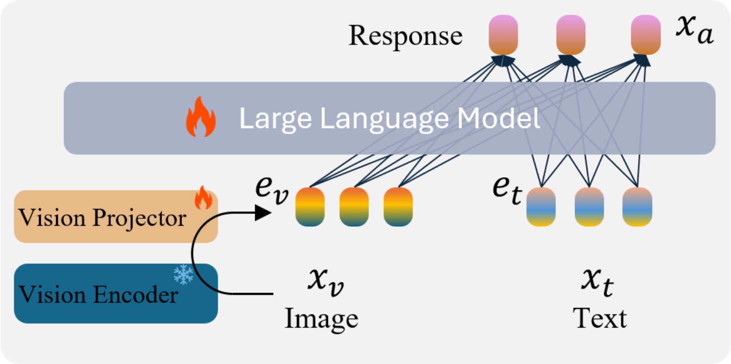

Figure 3: Model architecture.

To evaluate the effectiveness of our MCOUT framework, we conducted experiments using four diverse vision-language datasets: VQAv2 [24], MMMU [25], ScienceQA [26], and MMStar [27]. These datasets assess the model's reasoning capabilities across multimodal tasks, including VQA, scientific reasoning, and general knowledge understanding, with a focus on image-text integration. The VQAv2 dataset, used for pretraining, contains 443,757 question-answer pairs associated with images from the COCO dataset, emphasizing tasks like object recognition, attribute identification, and spatial reasoning.

The MMMU dataset, employed for fine-tuning, includes approximately 150 training samples and 900 validation samples. We also utilize the ScienceQA dataset, which focuses on scientific reasoning across natural science, social science, and language science. For this dataset, we use a subset of 6,218 training samples that contain both text and image contexts. The subset was chosen to preserve modality and format distributions while enabling fair ablations (MCOUT-Base/MCOUT-Multi, N_t, and μ) within a single-GPU training. The MMStar dataset, used exclusively for testing, consists of 1,500 test samples with curated image-question-answer triplets, designed for challenging visual reasoning tasks like object counting and scene understanding. All datasets are preprocessed to ensure compatibility with MCOUT's image-based pipeline, with images resized to 224×224 pixels and text tokenized to a maximum context length of 1024 tokens, interleaved with visual embeddings for unified input processing.

For training, we develop a multimodal model as described in Section 2.1, consisting of a pre-trained CLIP vision encoder and a Llama 3.2 1B language model. We pretrained the model on the VQAv2 training dataset for 1 epoch, followed by fine-tuning on ScienceQA and MMMU for 10 epochs. The model employs 8-bit precision, freezes the vision model, and uses LoRA (rank 64, alpha 16) for efficient adaptation. Training is conducted on a single CUDA device with 2 compute workers, using a batch size of 4 and a linear warmup cosine learning rate schedule (initial LR: 1×10^{-5}, minimum LR: 1×10^{-6}, warmup LR: 1×10^{-6}, weight decay: 0.05). The number of latent thoughts is experimented with values of 5 and 10 for both MCOUT-Base and MCOUT-Multi approaches, enabling iterative reasoning in a continuous latent space. During inference, we set the temperature to 0.1 for all experiments.

B Benchmarking

Table 2: Performance on the ScienceQA test set.

Models	Parameters (B)	accuracy (%)	BLEU
Our experiments			
Baseline	1	56.17	51.48
MCOUT-Base (N_t = 5)	1	58.60 (↑ 4.33%)	52.44 (↑ 1.87%)
MCOUT-Multi (N_t = 5)	1	58.45 (↑ 4.05%)	**52.60 (↑ 2.18%)**
MCOUT-Base (N_t = 10)	1	**58.86 (↑ 4.79%)**	52.31 (↑ 1.61%)
MCOUT-Multi (N_t = 10)	1	58.20 (↑ 3.61%)	52.27 (↑ 1.53%)
Literature reports			
Kosmos2 [19]	1.7	32.70	–
SilVar [17]	7	63.21	–
LLaVA-7B [22]	7	41.10	–
InstructBLIP-7B [20]	8	54.10	–
OpenFlamingo [23]	9	44.80	–
Qwen-VL [21]	9.6	61.10	–
MiniGPT-4 [31]	13	47.71	–
LLaMA2-13B [32]	13	55.78	–
LLaVA-13B [32]	13	47.74	–
PandaGPT-13B [33]	13	63.20	–

To evaluate the MCOUT framework, we compare MCOUT-Base and MCOUT-Multi against our baseline VLM without latent reasoning. Evaluations are conducted on the ScienceQA and MMMU validation sets and the MMStar test set, using accuracy and BLEU. We also compare our small VLM with other models. Tables 2, 3, and 4 summarize the results of our models on the ScienceQA, MMMU validation and MMStart benchmark, respectively.

For ScienceQA, as shown in Table 2, MCOUT-Base ($N_t = 10$) achieves the highest accuracy at 58.86% (up 4.79%), while MCOUT-Multi ($N_t = 5$) leads in BLEU at 52.60 (up 2.18%), excelling in image-heavy scientific reasoning due to its multimodal attention mechanism. With 1B parameters, both variants outperform larger models like Kosmos-2 (1.7B, 32.70%), LLaVA-7B/13B (41.10%–47.74%), and MiniGPT-4-13B (47.71%), and closely match InstructBLIP-7B (8B, 54.10%) and LLaMA-2-13B (55.78%), showcasing MCOUT's efficiency in leveraging iterative reasoning for robust performance.

For MMMU, as illustrated in Table 3, MCOUT-Base ($N_t = 5$) achieves the highest gains, with accuracy at 27.53% (up 8.21%) and BLEU at 27.54 (up 8.31%). MCOUT-Multi ($N_t = 10$) follows closely with 7.54% and 7.58% gains in accuracy and BLEU, respectively, leveraging multimodal attention for cross-modal tasks. With 1B parameters, MCOUT outperforms Kosmos-2 and MiniGPT-4 variants, and nearly matches OpenFlamingo-9B and Qwen-VL, demonstrating strong efficiency in college-level reasoning.

For MMStar, as illustrated in Table 4, MCOUT-Base ($N_t = 10$) improves accuracy and BLEU by 3.98%, while MCOUT-Multi ($N_t = 10$) gains 3.74% in both metrics, enhancing fine-grained visual reasoning through iterative thought generation. Despite its 1B parameters, MCOUT outperforms Kosmos-2, MiniGPT-4-v1-7B, MiniGPT-4-v2, and PandaGPT-13B, and closely rivals OpenFlamingo-9B and LLaVA-7B, highlighting its efficiency in challenging visual tasks.

Table 3: Performance on the MMMU validation set.

Models	**Parameters (B)**	**accuracy (%)**	**BLEU**
Our experiments			
Baseline	1	25.44	25.44
MCOUT-Base (N_t = 5)	1	**27.53 (↑ 8.21%)**	**27.54 (↑ 8.31%)**
MCOUT-Multi (N_t = 5)	1	27.18 (↑ 6.79%)	27.19 (↑ 6.82%)
MCOUT-Base (N_t = 10)	1	27.52 (↑ 8.18%)	**27.54 (↑ 8.31%)**
MCOUT-Multi (N_t = 10)	1	27.36 (↑ 7.54%)	27.37 (↑ 7.58%)
Literature reports			
Kosmos 2 [19]	1.7	23.7	–
MiniGPT-4-v1-7B [31]	7	23.6	–
LLaVA-v1.5-7B [22]	7	33.7	–
MiniGPT-4-v2 [34]	7	25.0	–
OpenFlamingo v2 [23]	9	28.8	–
Qwen-VL [21]	9.6	29.6	–
LLaVA-v1.5-13B [22]	13	37.0	–
PandaGPT-13B [33]	13	32.9	–

Table 4: Performance on the MMStar test set.

Models	**Parameters (B)**	**accuracy (%)**	**BLEU**
Our experiments			
Baseline	1	25.13	25.14
MCOUT-Base (N_t = 10)	1	**26.13 (↑ 3.98%)**	**26.14 (↑ 3.98%)**
MCOUT-Multi (N_t = 10)	1	26.07 (↑ 3.74%)	26.08 (↑ 3.74%)
Literature reports			
Kosmos2 [19]	1.7	24.9	–
MiniGPT-4-v1-7B [31]	7	16.3	–
MiniGPT-4-v2 [34]	7	21.3	–
LLaVA-7B [22]	7	27.1	–
OpenFlamingo v2 [23]	9	26.9	–
Qwen-VL-Chat [21]	9.6	34.5	–
PandaGPT-13B [33]	13	25.6	–

C Auxiliary Loss Study

To investigate the impact of the auxiliary weight μ in the MCOUT loss function (Equation 9), we conduct an ablation study with the impact of the auxiliary weight μ in the MCOUT loss function with $N_t = 5$, as shown in Table 5. $\mu = 0.3$ yields the highest performance, improving ScienceQA accuracy by 4.33%, and MMMU accuracy by 8.23%, highlighting the importance of balancing auxiliary thought supervision for effective multimodal reasoning. Higher μ values (0.5, 0.8) reduce gains, suggesting overemphasis on intermediate thoughts may disrupt final output optimization, while $\mu = 0$ yields moderate improvements. Although using an auxiliary loss boosts model performance, it increases training time based on our experiments.

Table 5: Ablation study for ScienceQA test and MMMU val using N_t = 5.

Models	**Auxiliary**	**ScienceQA test**		**MMMU val**	
	weight (μ)	**accuracy**	**BLEU**	**accuracy**	**BLEU**
Baseline		56.17	51.48	25.44	25.44
MCOUT-Base	0	58.12 (↑ 3.47%)	52.05 (↑ 1.11%)	27.41 (↑ 7.75%)	27.43 (↑ 7.82%)
MCOUT-Base	0.3	**58.60 (↑ 4.33%)**	**52.44 (↑ 1.87%)**	**27.53 (↑ 8.23%)**	**27.54 (↑ 8.27%)**
MCOUT-Base	0.5	57.56 (↑ 2.48%)	52.10 (↑ 1.20%)	26.44 (↑ 3.93%)	26.44 (↑ 3.93%)
MCOUT-Base	0.8	57.52 (↑ 2.40%)	52.00 (↑ 1.01%)	25.90 (↑ 1.81%)	25.91 (↑ 1.85%)