Search-MM: Benchmarking Multimodal Agentic RAG with Structured Reasoning Chains

Anonymous Author(s)
Affiliation
Address
email

Abstract

Retrieval-Augmented Generation (RAG) has become a key paradigm for grounding multimodal large language models (MLLMs) in external evidence. Current MM-RAG benchmarks, however, emphasize simplified QA tasks with shallow reasoning depth, falling short in evaluating agentic RAG behaviors such as iterative planning and retrieval. We present Search-MM, a benchmark with golden, hop-wise reasoning chains that specify sub-questions, retrieval modalities, supporting facts, and intermediate answers, enabling fine-grained analysis of retrieval planning and reasoning accuracy. To ensure fidelity, we propose HAVE, a hop-wise verification procedure that filters hallucinated or redundant steps. Search-MM covers five representative reasoning structures and consists of 3,333 high-quality examples. We further develop an agentic MM-RAG pipeline and introduce three chain-level metrics to jointly assess answer accuracy and intermediate retrieval fidelity. Experiments benchmark MLLMs under this framework, revealing key challenges in modality-aware planning and the trade-off between retrieval effectiveness and efficiency.

1 Introduction

Retrieval-Augmented Generation (RAG) has emerged as a key paradigm for enabling large language models in external evidence [5, 19, 4]. Facing various data modalities, multimodal RAG (MM-RAG) is proposed and is expected to retrieve and integrate text and image evidence to support knowledge-intensive reasoning [17, 16, 1]. To boost the further development of MM-RAG, a few corresponding benchmark datasets have recently been proposed [6, 12, 11]. Although materializing the concept, those pioneering efforts are nascent with the simple question-answer format that does not need complex reasoning iterations [6], limited length of reasoning steps [12, 11], and reliance on costly and unstable online search [11]. The above shortcomings hinder the agentic RAG development in the multimodal research domain.

Compared with classic RAG, agentic RAG systems [15, 10, 7, 8, 20, 3, 14] often exhibit iterative task decomposition, re-verification, and evidence planning, aiming to adaptively decide when and what to retrieve during reasoning. To effectively evaluate the above tasks in realistic multimodal reasoning scenarios, at least, the benchmarks require *step-wise annotations* that specify both the sequence of sub-questions and the modality of each retrieval step, while also distinguishing different retrieval–reasoning *structures* (e.g., text-initiated versus image-initiated chains, or parallel multimodal forks). Designing such a benchmark is challenging and non-trivial, as it needs to align with long, adaptive retrieval workflows while capturing diverse reasoning patterns to support fine-grained error analysis.

Hence, in this paper, we first introduce **Search-MM**, a benchmark for **agentic MM-RAG** with *structured multi-hop reasoning chains*. In Search-MM, each question sample is associated with a golden and step-wise annotated trajectory that specifies the sub-question sequence, the modality of

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

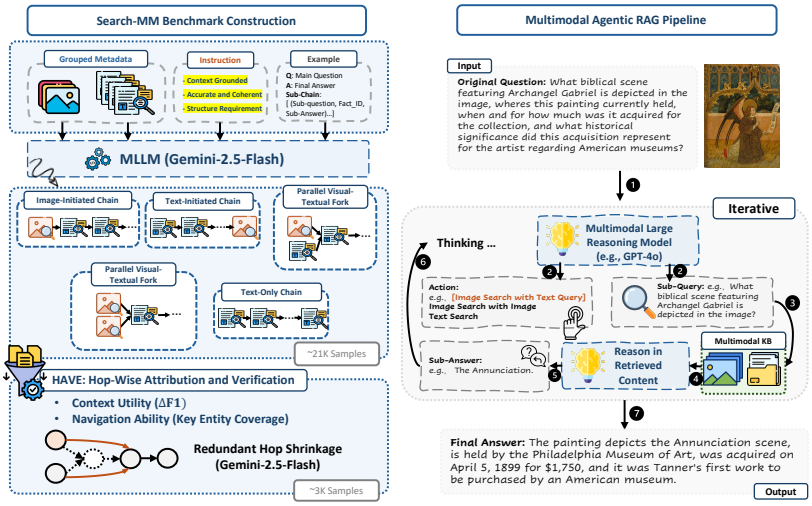

Figure 1: Overview of **Search-MM** benchmark and evaluation pipeline. **Left**: Benchmark construction with five reasoning structures and a hop-wise attribution and verification (HAVE) process to ensure retrieval necessity of each step. **Right**: Multimodal agentic RAG pipeline, where a reasoning model iteratively decomposes queries, retrieves multimodal evidence, and integrates it to generate the final answer.

each retrieval, the unique supporting fact, and the intermediate answer. This organization enables fine-grained attribution, chain-level evaluation, and lays the foundation for future process-level reward modeling in agentic multimodal reasoning.

In addition to multimodality, the reasoning chain in our Search-MM is also **long**, **category-diversified**, and **hop-non-redundant**. First, to reflect the diversity of real-world agentic MM-RAG cases, our Search-MM spans five representative reasoning structures, i.e., (i) *Text-Only Chain*, (ii) *Image-Initiated Chain*, (iii) *Text-Initiated Chain*, (iv) *Parallel Visual-Textual Fork*, and (v) *Multi-Image Fork*, capturing both serial and parallel reasoning patterns across modalities, as shown in Figure 1. Second, to ensure that each reasoning step (i.e., hop) in the chain is inference necessary and structurally meaningful, we propose **HAVE**, a Hop-Wise Attribution and Verification of Evidence procedure that filters out spurious or redundant hops via utility- and navigation-based diagnostics. This results in a high-quality benchmark of 3,333 well-annotated examples. Third, the average length of questions in Search-MM is 3.7 hops and leads the SOTA benchmarks [6, 12, 11].

For evaluation, beyond traditional answer-level accuracy, we introduce three **chain-level evaluation metrics**: (i) *LLM-as-a-Judge* for open-ended reasoning quality, (ii) *Structure-Aware Hit Rate* for per-step grounding fidelity, and (iii) *Rollout Deviation* to quantify execution drift. We conduct a **comprehensive evaluation** of three MLLM backbones, i.e., `GPT-4o-Mini`, `Gemini-2.5-Flash`, and `Gemini-2.5-Pro`, on the Search-MM benchmark, comparing their capabilities in retrieval planning and multimodal reasoning. For a fair comparison, we further develop **a unified agentic MM-RAG pipeline** that dynamically plans, retrieves, and fuses multimodal evidence conditioned on the evolving chain state. Our analysis also reveals how **over-retrieval** and **lack of retrieval** affect performance across different chain types, underscoring the need for better modality-aware planning and stopping criteria to balance retrieval effectiveness and efficiency.

2 Search-MM Benchmark

2.1 Search-enhanced Long Reasoning Chains

Structured Chain Typology Pattern and Data Preparation. To reflect the diversity of real-world agentic MM-RAG workflows, we identify five recurring search-enhanced reasoning structures and instantiate them as distinct chain types, each consisting of sub-questions, retrieval evidence, and intermediate answers (Figure 2). Specifically, (i) *Text-Only Chain* serves as a baseline for structured textual reasoning; (ii) *Image-Initiated Chain*, (iii) *Text-Initiated Chain*, and (iv) *Parallel Visual–Textual Fork* capture single-image settings with different initiation or branching strategies; and (v) *Multi-Images Fork* represents multi-image coordination. Data ex-

Figure 2: Distribution of five reasoning mechanisms in Search-MM, with outer segments showing hop-level diversity (2–5 hops showing).

Table 1: Evaluation of various models on the **Search-MM** benchmark when performing **agentic MM-RAG**. We report results across five reasoning graph categories. Best results are in **bold**, second-best are underlined.

Reasoning Graph	Model	Answer Accuracy			Chain Alignment		Golden F1 (↑)
		F1 (↑)	ΔF1 (↑)	LJ (↑)	HPS (↑)	RD (↓)	
Text-Only Chain	GPT-4o-Mini	30.96	30.94	2.58	**21.94**	1.13	61.90
	Gemini-2.5-Flash	34.03	33.96	2.49	20.59	**1.04**	**67.72**
	Gemini-2.5-Pro	**34.47**	**34.46**	**2.66**	21.59	1.07	62.42
Image-Initiated Chain	GPT-4o-Mini	36.49	34.18	2.63	27.51	1.46	68.29
	Gemini-2.5-Flash	44.10	37.38	3.01	**31.46**	2.91	**72.39**
	Gemini-2.5-Pro	**47.61**	**42.76**	**3.18**	25.90	**1.05**	69.83
Multi-Images Fork	GPT-4o-Mini	24.00	19.16	2.13	16.04	1.94	59.46
	Gemini-2.5-Flash	36.80	31.89	2.35	13.45	1.66	**64.40**
	Gemini-2.5-Pro	**40.37**	**36.58**	**2.76**	**18.68**	**1.40**	61.29
Parallel Visual-Textual Fork	GPT-4o-Mini	21.98	21.65	2.20	15.00	1.71	53.98
	Gemini-2.5-Flash	29.92	27.94	2.58	11.43	2.70	**57.99**
	Gemini-2.5-Pro	**34.83**	**34.19**	**2.99**	**16.74**	**1.29**	53.46
Text-Initiated Chain	GPT-4o-Mini	30.11	18.46	2.41	**27.18**	1.76	49.47
	Gemini-2.5-Flash	43.55	26.34	3.30	25.20	1.20	**66.27**
	Gemini-2.5-Pro	**45.30**	**29.89**	**3.62**	19.51	**0.95**	55.94

amples for each structure are provided in Appendix B.

We construct Search-MM by clustering Wikipedia entities [2] into topical neighborhoods and prompting `Gemini-2.5-Flash` to generate structured multi-hop questions aligned with the five mechanisms. To guarantee necessity, we filter the knowledge base to remove entries that could also answer sub-questions, ensuring each chain follows a unique, meaningful trajectory.

Quality Verification. When constructing Search-MM, we found that multimodal LLMs often produce hallucinated or redundant reasoning steps that hinder faithful evaluation. To address this, we propose HAVE (Hop-wise Attribution and Verification Evaluation), a filtering mechanism that verifies the necessity of each step. We measure context utility by removing steps and checking the drop in answer F1, and assess navigation utility by testing whether key entities are carried forward into later sub-questions. For borderline cases, we apply `Gemini-2.5-Flash` for chain shrinkage and re-validation. Dataset composition and statistics are shown in Figure 2 and Appendix C.

2.2 Evaluation

Evaluation Metrics. To comprehensively evaluate both answer correctness and reasoning quality, we designed four complementary metrics. First, we compute the standard token-level answer by **F1**. We also report **ΔF1**, the gain of agentic MM-RAG over the same model question-answering without context retrieval, and **Golden F1**, the upper bound when providing gold reasoning chains and retrieval content. To capture semantic correctness beyond surface-level matches, we employ an **LLM-as-a-Judge (LJ)** approach, where a strong reasoning model (`Gemini-2.5-Pro`) assesses the generated reasoning chain against the gold standard based on *accuracy*, *coherence*, *knowledge entity coverage*, and *step alignment*.

Then, to evaluate step-wise retrieval accuracy, we introduce **Hit per Step (HPS)**. Let the predicted and golden reasoning chains be sets of steps $C_p = \{p_1, \ldots, p_m\}$ and $C_g = \{g_1, \ldots, g_n\}$, respectively. We first construct a bipartite graph between C_p and C_g, where edge weights are defined by the semantic similarity of the evidence retrieved for each pair (p_i, g_j). After solving for the maximum-weight matching M^*, we assess the retrieval hit of these matched pairs. A golden step $g_j \in C_g$ is considered a 'hit', if it is matched with a predicted step p_i (i.e., $(p_i, g_j) \in M^*$) and their underlying evidence sets are identical. HPS is then the fraction of golden steps that were correctly matched and replicated: $\text{HPS} = |C_g|^{-1} \sum_{g_j \in C_g} \mathbb{I}\left(\exists p_i \text{ s.t. } (p_i, g_j) \in M^* \wedge \text{evidence}(p_i) = \text{evidence}(g_j)\right)$, where $\mathbb{I}(\cdot)$ is the indicator function. Finally, to quantify the structural difference in reasoning length, we measure **Rollout Deviation (RD)** as the absolute difference in the number of steps: $\text{RD} = ||C_p| - |C_g||$.

Evaluation Setting. We propose a multimodal agentic RAG pipeline for evaluating adaptive retrieval on the structured, step-wise reasoning chains in Search-MM, drawing inspiration from recent advances in agent-based retrieval planning [10]. As shown in Figure 1, the pipeline iteratively follows a decompose—retrieve—synthesize process to plan sub-queries, gather multimodal evidence, and generate the final answer.

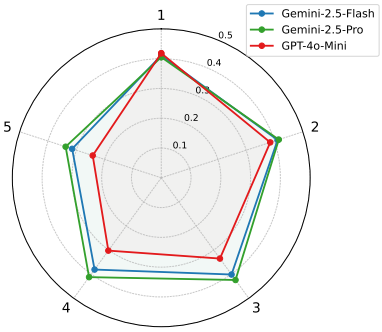

Figure 3: Model F1 across samples with different chain lengths, showing a consistent drop in performance as the chain length increases.

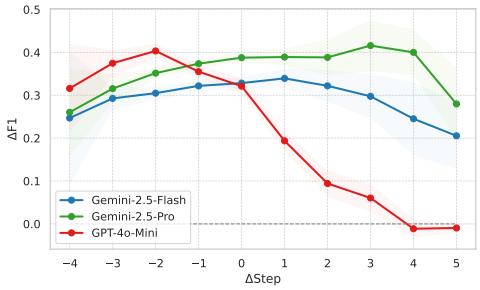

Figure 4: Model ΔF1 (difference in F1 between our generated response with context and without context) vs. Δ steps (difference in length between generated and golden reasoning chains), where larger positive Δ steps indicate more over-retrieval.

3 Experiments

Comparing results. Table 1 and Appendix D.2 show that retrieval quality (HPS) highly aligns with answer accuracy in most cases, confirming that many failures are due to sub-questions failing to retrieve the correct evidence. `Gemini-2.5-Pro` outperforms the others across metrics, with a notably smaller gap to the golden F1, indicating stronger retrieval planning and reasoning. Among the chain types, *Multi-Image Fork* and *Parallel Visual-Textual Fork* are the most challenging because they require coordinating information from multiple modalities. In experiments, models often default to text retrieval rather than conducting sufficient image search and grounding their reasoning in visual content and associated captions. This suggests limitations in current models' ability to plan modality-specific queries and highlights the need for better multimodal planning.

Performance vs. Chain Length. We observe a consistent performance drop as the length of the reasoning chain increases, see Figure 3. While all models are affected, `Gemini-2.5-Pro` exhibits the smallest decline, demonstrating stronger robustness in multi-hop planning and evidence integration. In contrast, `GPT-4o-Mini` shows a more rapid degradation, perhaps resulting from its weaker retrieval coordination and limited context tracking capacity. These results highlight the compounding difficulty of longer chains, where errors in earlier steps can cascade and undermine subsequent reasoning.

Over-Retrieval Analysis. We analyze the deviation between model-generated retrieval turns and the golden reasoning chain, focusing on how over-retrieval and under-retrieval affect performance. For both `Gemini-2.5-Flash` and `Pro`, a small degree of over-retrieval (Δ Step = 1 or 2) improves answer accuracy, as it may help compensate for imperfect planning or missed evidence. However, excessive over-retrieval leads to sharp performance degradation when the model pursuing increasingly irrelevant directions when failing to retrieve useful context. In contrast, `GPT-4o-Mini` benefits most from shorter chains and suffers more from over-retrieval: its limited reasoning and integration capacity may lead to context confusion and an inability to fuse multiple pieces of retrieved content, resulting in steeper decline as the chain lengthens unnecessarily. These trends highlight the importance of precise retrieval planning and stopping criteria in agentic MM-RAG systems to balance retrieval efficiency and reasoning effectiveness.

4 Conclusion.

We introduce Search-MM, a benchmark for structured, step-wise multimodal retrieval-augmented reasoning, covering five diverse reasoning mechanisms. Through fine-grained annotations, hop-wise attribution and verification, and the design of new chain-level metrics, Search-MM enables rigorous diagnosis of retrieval planning and reasoning quality. Our analyses underscore the need for adaptive, modality-aware search strategies in agentic MM-RAG systems. In future work, we will broaden evaluations to include more state-of-the-art reasoning models. We envision Search-MM as a foundation for developing interpretable, robust, and efficient multimodal agents, and for advancing process-level reward modeling in multimodal reasoning. Beyond evaluation, we plan to expand the benchmark to new domains, and we hope it will foster community efforts toward principled evaluation standards for agentic multimodal reasoning.

References

[1] Mohammad Mahdi Abootorabi, Amirhosein Zobeiri, Mahdi Dehghani, Mohammadali Mohammadkhani, Bardia Mohammadi, Omid Ghahroodi, Mahdieh Soleymani Baghshah, and Ehsaneddin Asgari. Ask in any modality: A comprehensive survey on multimodal retrieval-augmented generation. In *Annual Meeting of the Association for Computational Linguistics*, 2025.

[2] Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. Webqa: Multihop and multimodal qa. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16495–16504, 2022.

[3] Mingyang Chen, Tianpeng Li, Haoze Sun, Yijie Zhou, Chenzheng Zhu, Haofen Wang, Jeff Z Pan, Wen Zhang, Huajun Chen, Fan Yang, et al. Learning to reason with search for llms via reinforcement learning. *arXiv preprint arXiv:2503.19470*, 2025.

[4] Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. A survey on RAG meeting llms: Towards retrieval-augmented large language models. In Ricardo Baeza-Yates and Francesco Bonchi, editors, *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024, Barcelona, Spain, August 25-29, 2024*, pages 6491–6501. ACM, 2024.

[5] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *CoRR*, abs/2312.10997, 2023.

[6] Wenbo Hu, Jia-Chen Gu, Zi-Yi Dou, Mohsen Fayyaz, Pan Lu, Kai-Wei Chang, and Nanyun Peng. Mrag-bench: Vision-centric evaluation for retrieval-augmented multimodal models. In *The Thirteenth International Conference on Learning Representations*.

[7] Jerry Huang, Siddarth Madala, Risham Sidhu, Cheng Niu, Hao Peng, Julia Hockenmaier, and Tong Zhang. Rag-rl: Advancing retrieval-augmented generation via rl and curriculum learning. *arXiv preprint arXiv:2503.12759*, 2025.

[8] Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*, 2025.

[9] Meng-Chieh Lee, Qi Zhu, Costas Mavromatis, Zhen Han, Soji Adeshina, Vassilis N Ioannidis, Huzefa Rangwala, and Christos Faloutsos. Agent-g: An agentic framework for graph retrieval augmented generation.

[10] Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. Search-o1: Agentic search-enhanced large reasoning models. *arXiv preprint arXiv:2501.05366*, 2025.

[11] Yangning Li, Yinghui Li, Xinyu Wang, Yong Jiang, Zhen Zhang, Xinran Zheng, Hui Wang, Hai-Tao Zheng, Fei Huang, Jingren Zhou, et al. Benchmarking multimodal retrieval augmented generation with dynamic vqa dataset and self-adaptive planning agent. In *The Thirteenth International Conference on Learning Representations*.

[12] Zhenghao Liu, Xingsheng Zhu, Tianshuo Zhou, Xinyi Zhang, Xiaoyuan Yi, Yukun Yan, Yu Gu, Ge Yu, and Maosong Sun. Benchmarking retrieval-augmented generation in multi-modal contexts. *CoRR*, abs/2502.17297, 2025.

[13] Zhili Shen, Chenxin Diao, Pavlos Vougiouklis, Pascual Merita, Shriram Piramanayagam, Enting Chen, Damien Graux, Andre Melo, Ruofei Lai, Zeren Jiang, et al. Gear: Graph-enhanced agent for retrieval-augmented generation. *arXiv preprint arXiv:2412.18431*, 2024.

[14] Zhongxiang Sun, Qipeng Wang, Weijie Yu, Xiaoxue Zang, Kai Zheng, Jun Xu, Xiao Zhang, Yang Song, and Han Li. Rearter: Retrieval-augmented reasoning with trustworthy process rewarding. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1251–1261, 2025.

[15] Liang Wang, Haonan Chen, Nan Yang, Xiaolong Huang, Zhicheng Dou, and Furu Wei. Chain-of-retrieval augmented generation. *ArXiv*, abs/2501.14342, 2025.

[16] Peng Xia, Peng Xia, Kangyu Zhu, Haoran Li, Haoran Li, Weijia Shi, Sheng Wang, Linjun Zhang, James Zou, and Huaxiu Yao. Mmed-rag: Versatile multimodal rag system for medical vision language models. *ArXiv*, abs/2410.13085, 2024.

[17] Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Rich James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. Retrieval-augmented multimodal language modeling. *arXiv preprint arXiv:2211.12561*, 2022.

[18] Han Zhang, Langshi Zhou, and Hanfang Yang. Learning to retrieve and reason on knowledge graph through active self-reflection. *arXiv preprint arXiv:2502.14932*, 2025.

[19] Penghao Zhao, Hailin Zhang, Qinhan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang, Wentao Zhang, Jie Jiang, and Bin Cui. Retrieval-augmented generation for ai-generated content: A survey. *arXiv preprint arXiv:2402.19473*, 2024.

[20] Yuxiang Zheng, Dayuan Fu, Xiangkun Hu, Xiaojie Cai, Lyumanshan Ye, Pengrui Lu, and Pengfei Liu. Deepresearcher: Scaling deep research via reinforcement learning in real-world environments. *arXiv preprint arXiv:2504.03160*, 2025.

Appendix

Contents

A Related Work

We review three relevant lines of research. First, *multimodal RAG* extends retrieval beyond text to images and heterogeneous sources, but most benchmarks still rely on fixed-step designs. Second, *agentic RAG* treats retrieval as an adaptive, sequential process, yet multimodal variants remain scarce. Finally, *structure-aware benchmarks* supervise intermediate reasoning, motivating our Search-MM benchmark that unifies multimodality, agentic behaviors, and step-wise structural evaluation.

A.1 Multimodal RAG

Early retrieval-augmented generation focused on textual retrievers for LMs [17, 19]. More recently, RAG has been extended to multimodal settings by incorporating vision-language encoders, multimodal retrievers, and MLLMs [16, 1, 6]. While these works demonstrate strong performance in open-domain QA or medical domains, most existing multimodal RAG benchmarks adopt *fixed-step* settings, e.g., constraining the model to a predetermined number of text or image retrievals to answer a question. Such rigid designs prevent evaluating whether the agent can adaptively decide retrieval steps, and make it difficult to attribute errors to planning, modality selection, or evidence fusion.

A.2 Agentic RAG

Agentic RAG pipelines reformulate retrieval as a sequential decision-making process, where models iteratively decompose tasks, trigger retrieval, and integrate evidence during reasoning [7]. Recent approaches introduce explicit `<Search>` actions and train agents to plan, verify, and re-use evidence [8, 20], yielding more interpretable trajectories [3, 14]. Other work analyzes retrieval chains and highlights failure modes such as over- or lack-retrieval [15, 10]. However, most existing evaluations remain limited to the textual domain. Multimodal agentic RAG variant [11]s are still scarce, and current datasets do not provide *golden intermediate chains*, which hinders diagnosis of adaptive retrieval behaviors across modalities.

A.3 Structure-Aware Benchmarks for Agentic RAG

A related line of work investigates structured agents operating over symbolic KBs or graphs, often with schema-constrained queries or specialized tools [9, 13, 18]. These improve symbolic grounding but mainly focus on *what* to retrieve. In contrast, benchmarking agentic MM-RAG requires supervision over the *structure of the retrieval chain* itself—covering initiation modality, hop order, and parallel vs. serial patterns. To our knowledge, no prior multimodal dataset provides such structured, step-wise annotations. Our Search-MM fills this gap by supplying golden reasoning chains with modality-specific retrieval steps, enabling fine-grained attribution, analysis of over-/under-retrieval, and evaluation of diverse multimodal reasoning structures.

B Data Example

B.1 Image-Initiated Chain

Example Reasoning Chain — Image-Initiated Chain

Caption: Archangel Gabriel from an Annunciation - Google Art Project

Main Question. What biblical scene featuring Archangel Gabriel is depicted in the image, where is this painting currently held, when and for how much was it acquired for the collection, and what historical significance did this acquisition represent for the artist regarding American museums?

Final Answer. The painting depicts the *Annunciation* scene, is held by the *Philadelphia Museum of Art*, was acquired on *April 5, 1899* for *$1,750*, and it was *Tanner's first work to be purchased by an American museum*.

Reasoning Chain (4 steps):

1. **(Image)** What biblical scene featuring Archangel Gabriel is depicted in the image?
   Image Evidence ID: `30332773`
   Answer: The Annunciation.
2. **(Text)** Which museum holds this painting?
   Text Evidence ID:
   `d5be2ae00dba11ecb1e81171463288e9_0`
   Answer: Philadelphia Museum of Art.
3. **(Text)** When and for what price was the painting acquired for the collection?
   Text Evidence ID:
   `d5be2ae00dba11ecb1e81171463288e9_1`
   Answer: April 5, 1899 for $1,750.
4. **(Text)** What was the historical significance of this acquisition for the artist regarding American museums?
   Text Evidence ID:
   `d5be2ae00dba11ecb1e81171463288e9_8`
   Answer: It was Tanner's first work to be purchased by an American museum.

B.2 Text-Initiated Chain

Example Reasoning Chain — Text-Initiated Chain

Caption: Solvang California Windmill.

Main Question. Does the image visually confirm Solvang's unique cultural character as described in the text?

Final Answer. *Yes*, the image prominently features a large windmill and buildings with distinct European/-Danish architectural styles, aligning with the description of Solvang as "a bit of Denmark" in California.

Reasoning Chain (2 steps):

1. **(Text)** How is Solvang culturally characterized according to the text?
   Text Evidence ID: `d5bd142a0dba11ecb1e81171463288e9_7`
   Answer: Solvang is described as "a bit of Denmark tucked into this beautiful California."
2. **(Image)** Does the image visually confirm this cultural characterization, and if so, how?
   Image Evidence ID: `30326437`
   Answer: Yes, the image prominently displays a large windmill and buildings with distinct European/Danish architectural styles, confirming its cultural description.

B.3 Text Chain

Example Reasoning Chain — Text Chain

Main Question. What are the key attributes of the Bentley Mulsanne, including its manufacturing period, engine specifications, origin of its name, and notable special editions?

Final Answer. The Bentley Mulsanne, a full-size luxury car *manufactured from 2010 to 2020*, is named after the Mulsanne Corner of the Le Mans racing circuit. It uses a *6.75 L Bentley L Series V8 engine*, and a notable special edition is the *"W.O. Edition"*, which features a piece of W.O. Bentley's personal car crankshaft.

Reasoning Chain (3 steps):

1. **(Text)** What type of vehicle is the Bentley Mulsanne, when was it manufactured, and what is the origin of its name?
   Text Evidence ID: `d5bd6ace0dba11ecb1e81171463288e9_15`
   Answer: The Bentley Mulsanne is a full-size luxury car that was manufactured from 2010 to 2020 and is named after the Mulsanne Corner of the Le Mans racing circuit.
2. **(Text)** What are the engine specifications of the Bentley Mulsanne?
   Text Evidence ID: `d5bd6ace0dba11ecb1e81171463288e9_6`
   Answer: The Mulsanne uses a 6.75 L (6,750 cc/411 in^3) Bentley L Series V8 engine, modified to meet Euro V emissions regulations.
3. **(Text)** What notable special edition of the Mulsanne was introduced and what was its unique feature?
   Text Evidence ID: `d5bd6ace0dba11ecb1e81171463288e9_11`
   Answer: The Mulsanne "W.O. Edition" was presented, featuring a piece of the crankshaft from W.O. Bentley's personal car displayed in the arm rest.

B.4 Parallel Visual-Textual Fork

Example Reasoning Chain — Parallel Visual-Textual Fork

Caption: Mandolin1.

Main Question. Given the appearance of the mandolin shown, how did Pasquale Vinaccia's 1835 innovation involving string material lead to structural changes in the instrument and its lasting impact on mandolin design?

Final Answer. Pasquale Vinaccia's 1835 innovation of using *steel wire strings*, which are visible on the mandolin, necessitated *strengthening the body* and *deepening the bowl* for increased *resonance*, and this method of stringing ultimately *became the dominant way* for mandolins.

Reasoning Chain (4 steps):

1. **(Image)** What type of strings are visible on the mandolin depicted, and how does this visual detail relate to Pasquale Vinaccia's historical improvements to the instrument?
   Image Evidence ID: `30204742`
   Answer: The mandolin prominently displays wire/steel strings."
2. **(Text)** What structural modifications were subsequently required for the mandolin's body due to the adoption of these new wire strings?
   Text Evidence ID: `d5be68020dba11ecb1e81171463288e9_15`
   Answer: The wire strings necessitated strengthening the body and deepening the bowl.
3. **(Text)** What specific acoustic quality was enhanced by deepening the mandolin's bowl as a consequence of these structural changes?
   Text Evidence ID: `d5be68020dba11ecb1e81171463288e9_3`
   Answer: Deepening the bowl increased tonal resonance.
4. **(Text)** Considering these improvements, what was the long-term impact of Vinaccia's decision to use steel strings on mandolin design and construction?
   Text Evidence ID: `d5be68020dba11ecb1e81171463288e9_5`
   Answer: His steel-stringing approach became the dominant way of stringing mandolins.

B.5 Multi-Images Fork

Example Reasoning Chain — Multi-Images Fork

Caption: Canada Pavilion

Caption: Morocco Pavilion

Main Question. Which of the depicted Epcot pavilions, the Canada or Morocco, features a prominent tall, slender tower as its main architectural centerpiece, and what is its historical significance regarding its addition to the World Showcase?

Final Answer. The Morocco Pavilion features *a prominent tall, slender tower*, and it was historically significant as *the first expansion pavilion added to World Showcase*, opening on September 7, 1984.

Reasoning Chain (4 steps):

1. **(Image)** What is the most prominent feature of the roof structure shown on the main building, the Canada Pavilion?
   Image Evidence ID: `30383576`
   Answer: A very steep, multi-tiered green roof with pointed spires.
2. **(Image)** What is the most prominent architectural feature defining the skyline of the building in the Morocco Pavilion?
   Image Evidence ID: `30021436`
   Answer: A tall, rectangular tower topped with a small dome.
3. **(Image)** Considering the contrast in primary vertical architectural elements between the Canada Pavilion's steep roof and the Morocco Pavilion's tower, which of these two pavilions would be best described as featuring a prominent tower rather than a very steep roof?
   Image Evidence ID: `30021436`
   Answer: The Morocco Pavilion.
4. **(Text)** According to the provided text, what significant historical fact is associated with the Morocco Pavilion regarding its status as a World Showcase addition?
   Text Evidence ID: `d5bef66e0dba11ecb1e81171463288e9_7`
   Answer: It was the first expansion pavilion to be added to World Showcase, opening on September 7, 1984.

C Dataset Statistics

Table 2: Statistics of Search-MM benchmark.

Statistic	**Number**
Parallel Visual-Textual Fork Samples	680
Image-Initiated Chain Samples	1,306
Text Chain Samples	945
Text-Initiated Chain Samples	169
Multi-Images Fork Samples	233
Total Samples	3,333
Total Images in the Knowledge Base	389,750
Total Documents in the Knowledge Base	784,473
Average Chain Length in Parallel Visual-Textual Fork	4.01
Average Chain Length in Image-Initiated Chain	3.82
Average Chain Length in Text Chain	3.65
Average Chain Length in Text-Initiated Chain	3.17
Average Chain Length in Multi-Images Fork	3.97
Sub-Chains with Text Retrieval	10,063 (79.72%)
Sub-Chains with Image Retrieval	2,550 (20.20%)

D Agentic MM-RAG Case Study

D.1 Success Case

Figure 5: Success case. The reasoning chain produced by the agent aligned with the golden reasoning chain and the final answer includes all key knowledge entities.

D.2 Failure Case

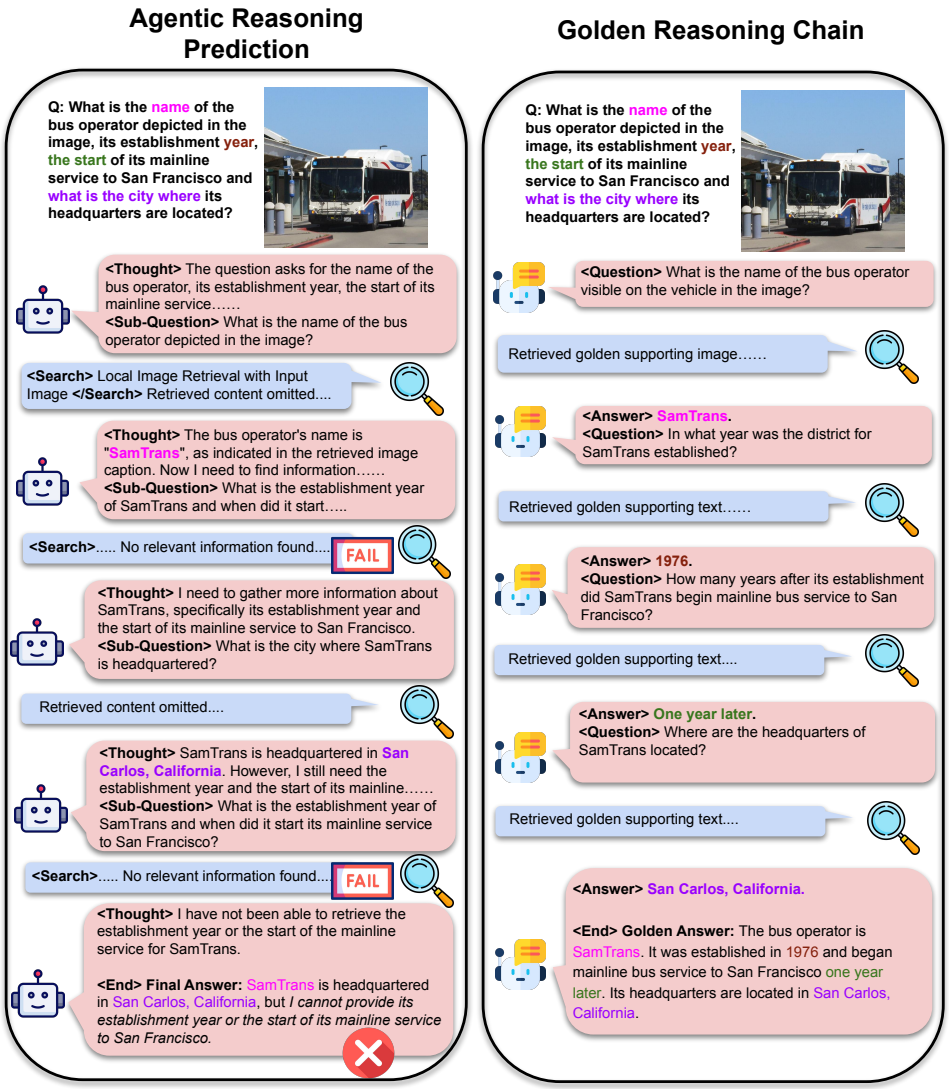

Figure 6: Failure case. While the agent successfully retrieves the first and the last hop information, it fails to retrieve the second and third hop information. Thus, the final answer does not include all key knowledge entities.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction provide overall explanation of Search-MM

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We provide our future work and limitation shortly in the conclusion.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.

- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [No]

   Justification: We do not include theoretical results.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have provided all the necessary details for reproduction.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation,

it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We will public our data and code upon acceptance.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide all the details in the experiment part.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Our method is evaluated across multiple data types and through extensive analyses and comparisons. We believe we don't need error bars in this paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The necessary computational resources are provided in the appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We carefully read the Code of Ethics and follow every instruction.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We will provide the Impact Statement in the appendix in the camera-ready version.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our algorithm doesn't involve any moral or safety considerations.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite all the papers we need.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.

- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We introduce new dataset in this paper and will upload the dataset to an anonymous link.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: We don't employ anyone.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: We don't have such potential risks.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: We only use LLMs to polish paper writing.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.