Can Frontier LLMs Match Natively Multimodal Embeddings? A Comparison on Hard-Negative Text-to-Image Retrieval

Anonymous ACL submission

Abstract

Multimodal retrieval and classification across different types of media, spanning text, images, video and audio, has traditionally relied on dual-encoder models that align visual and textual representations through contrastive learning. The March 2026 release of Gemini Embedding 2, Google's first natively multimodal embedding model to map text, images, video, audio, and documents into a single shared space, raises competition among multimodal retrieval systems. Simultaneously, frontier Large language models (LLMs) have also demonstrated strong visual understanding, raising the question of whether they can serve as effective zero-shot rankers. Our study provides the first direct comparison of native multimodal embeddings against LLM-based visual ranking on Flickr30k. We observe that GPT-4.1 and Claude Sonnet 4.6 perform on par with Gemini Embedding 2. Additionally, once embeddings are precomputed, multimodal embeddings are better suited for low-latency applications.

1 Introduction

Text-image retrieval, ranking a gallery of images by relevance to a natural language query, is a long-standing task with applications in visual search, content recommendation, and retrieval-augmented generation. There have been two recent developments in this space. First, LLMs with strong vision capabilities, such as GPT-4.1 (OpenAI, 2025) and Claude (Anthropic, 2024), can now receive images and text jointly, enabling a fundamentally different retrieval paradigm: presenting all candidate images alongside the query in one prompt and ranking them directly. Second, natively multimodal embedding models have recently become publicly available: Gemini Embedding 2 (GE2) (Google DeepMind, 2026) and Amazon Nova 2 (Nova 2) (Amazon, 2025), released in March, 2026 and October 2025, respectively, both project multiple modalities through a single shared backbone into a unified embedding space, unlike dual-encoder models that align independently encoded representations. The comparison of native multimodal embedding (GE2, Nova 2) with direct LLM visual ranking (GPT-4.1, Claude Sonnet 4.6) is timely and currently uncharacterized in the literature, specifically on (1) **Retrieval Accuracy** and (2) **Ranking Time** Our contributions are as follows:

- The first comparison of natively multimodal embedding models against frontier LLMs as zero-shot direct visual rankers on Flickr30k (Young et al., 2014), using semantically challenging hard-negative candidate sets without task-specific fine-tuning or intermediate caption generation.
- An empirical finding that GE2, GPT-4.1, and Claude Sonnet 4.6 achieve statistically indistinguishable retrieval accuracy, and embedding models rank 1000 queries in under two seconds (with pre-computation) compared to over three hours for LLMs.

2 Related Work

Dual-encoder cross-modal retrieval. CLIP (Radford et al., 2021) established the paradigm of learning aligned image and text encoders via large-scale contrastive pretraining, achieving strong zero-shot transfer on image-text retrieval benchmarks. ALIGN (Jia et al., 2021) demonstrated that data scale can compensate for annotation quality. More recent work, including BLIP-2 (Li et al., 2023) and EVA-CLIP (Fang et al., 2023), has extended these ideas with larger architectures and richer pretraining objectives, achieving state-of-the-art performance on COCO (Lin et al., 2014) and Flickr30k.
LLMs for retrieval and re-ranking. Sun et al. (2023) found that ChatGPT achieves competitive performance as a zero-shot document ranker on text-only benchmarks. Ma et al. (2023) showed that prompting LLMs to produce listwise rankings

outperforms pointwise scoring in text retrieval. In the multimodal setting, BLIP-2 (Li et al., 2023) and InstructBLIP (Dai et al., 2023) use LLMs to generate textual image descriptions, which are then matched against queries by a text retrieval system. Our work departs from this indirect approach by prompting LLMs to rank all candidate images together. MM-Embed (Jiang et al., 2024) fine-tunes a multimodal LLM-based retriever but focuses on learning retrieval representations rather than evaluating frontier models as zero-shot rankers. RagVL (Zhu et al., 2024) demonstrates that multimodal LLMs are strong rerankers in retrieval-augmented generation settings, though in a document QA context rather than image retrieval.

Natively multimodal embeddings. ImageBind (Girdhar et al., 2023) demonstrated joint embedding of six modalities using image-paired data as a binding signal, though its underlying encoders remain modality-specific. VLM2Vec (Jiang et al., 2025) takes a step further by fine-tuning a vision-language model as a unified embedding backbone, leveraging deep integration of vision and language features within a single transformer architecture rather than late fusion of independently encoded representations.

Hard negative mining. VSE++ (Faghri et al., 2018) introduced hard negative mining as a key training-time improvement, showing that selecting the most violating negatives substantially improves rank-based metrics. Robinson et al. (2021) also demonstrate that semantically hard negatives substantially change task difficulty.

3 Methodology

3.1 Dataset, Sampling and Evaluated Systems

Flickr30k (Young et al., 2014) provides 31,000 images each annotated with five independent captions. Although full-gallery retrieval on Flickr30k is largely saturated at the top of the leaderboard, our hard-negative candidate sets create a non-trivial task. We sample a pool of 5,000 images from Flickr30k, each paired with five independent human-written captions. To obtain a single representative query caption per image while controlling for annotator-style variance, we apply a *stratified* caption sampling: the 5,000 images are partitioned into five equally-sized groups of 1,000, each assigned exclusively to one caption index (0 to 4). The ground truth for every query is the image from which its caption was originally drawn.

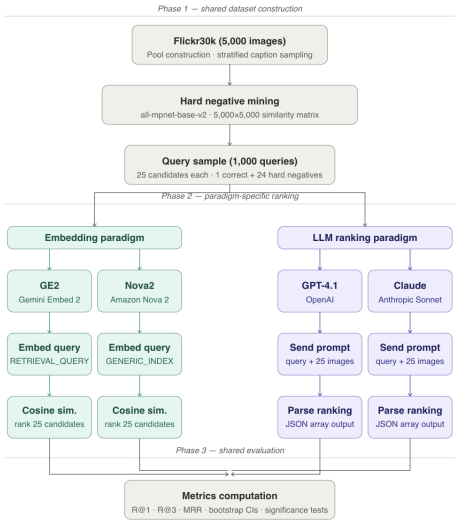


Figure 1: Overview of the three-phase evaluation pipeline: hard negative dataset construction, ranking via embedding models and LLMs, and evaluation.

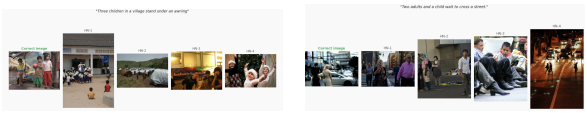


Figure 2: Examples of hard negative datasets showing candidate images for a given query.

3.2 Hard Negative Candidate Set Construction

For each of the 5,000 pool images, we constructed a candidate set of 25 images: the ground-truth image and the 24 hardest negatives. This candidate set size relatively balances task difficulty with the practical constraint of fitting all images within a single LLM prompt without exceeding context limits. Hard negatives are selected by embedding all representative captions with `sentence-transformers/all-mpnet-base-v2` (Reimers and Gurevych, 2019), a general-purpose text encoder that is intentionally selected to be decoupled from all evaluated retrieval systems. We then compute a 5,000×5,000 pairwise cosine similarity matrix, and retain the 24 most similar captions to the query (excluding the query itself). The images corresponding to these captions form the hard negative set. From the full pool, we randomly sample 1,000 captions as queries. Figure 1 shows the pipeline diagram.

We construct hard negatives in *caption* embedding space rather than image embedding space for two reasons. First, grounding distractor difficulty in caption-space similarity increases ambiguity, mak-

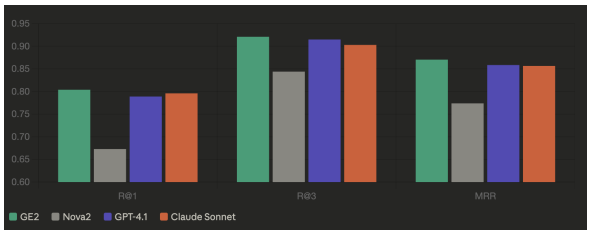


Figure 3: Evaluating the four systems on Recall@1, Recall@3 and MRR.

ing images hard to distinguish *given the caption*, which is precisely the retrieval difficulty we wish to evaluate. Second, since human perception of image similarity is naturally expressed through language, using caption similarity to construct candidate sets is a well-motivated design choice (details mentioned in Limitations). We note that constructing hard negatives in caption space may favor embedding-based retrieval methods. All evaluated systems (Gemini Embedding 2, Amazon Nova 2 Multimodal Embedding, GPT-4.1, and Claude Sonnet 4.6) are exposed to identical queries and candidate sets. Details about system configurations are provided in Appendix A

3.3 Evaluation Metrics and Statistical Tests

We report three metrics over 1,000 queries:

- **Recall@1 (R@1)**: Fraction of queries where the ground-truth image is ranked first.
- **Recall@3 (R@3)**: Fraction of queries where the ground-truth image appears in the top three.
- **Mean Reciprocal Rank (MRR)**: Mean of the reciprocal rank of the ground-truth image across all queries.

Additionally, we report 95% bootstrap confidence intervals for all three metrics (Appendix B) using 1,000 resamples (Efron and Tibshirani, 1993). For pairwise significance testing of R@1 (a binary outcome), we apply McNemar's test (McNemar, 1947). For MRR (a continuous outcome), we apply the Wilcoxon signed-rank test (Wilcoxon, 1945). We use a significance threshold of $\alpha = 0.05$ with Bonferroni correction for three pairwise comparisons.

4 Results and Discussion

Figure 3 visualizes the metric comparison. GE2 achieves the highest score across all three metrics, followed closely by Claude Sonnet 4.6 and GPT-4.1, with Amazon Nova 2 trailing by 13 percentage points. However, pairwise significance testing (Table 1) reveals a clear separation: Gemini Embedding 2, GPT-4.1, and Claude Sonnet 4.6 are statistically indistinguishable from one another on both R@1 (McNemar, all $p > 0.25$) and MRR (Wilcoxon, all $p > 0.09$), while the performance of Amazon Nova 2 is significantly different from every other system ($p < 0.0001$ on both tests). The absence of a statistically significant gap between the embedding paradigm (Gemini Embedding 2) and the LLM reasoning paradigm (GPT-4.1, Claude Sonnet 4.6) is the primary finding of this study (given the dataset and candidate set). Dedicated multimodal embedding and joint visual reasoning achieve competitive retrieval accuracy on hard-negative candidates drawn from a large image pool.

GE2's accuracy parity with LLMs-based rankers suggests that its embedding space is optimized for retrieval, whereas LLMs are generalist models that rank through prompting. The accuracy parity could also be potentially attributed to embedding models' similarity scores computed independently for each query-candidate pair, making them robust to candidate set composition and free from position or distractor biases (Wang et al., 2023). However, we minimized this positioning bias in LLMs by randomizing the order per query. In contrast, LLMs have an advantage that they receive all 25 images simultaneously and can reason about relative visual content across candidates before producing a ranking. This cross-candidate reasoning might be an advantage for the LLMs. As a qualitative example, in Figure 4 we observe that GE2 retrieved the correct image despite it being dark, while Nova 2 and GPT-4.1 selected the same image as their top-ranked image. Additional examples are provided in Appendix F.

Amazon Nova 2 trails GE2 by 13.1 percentage points on Recall@1. We attribute this gap to the use of `GENERIC_INDEX` as the embedding purpose, which is designed for general vector database creation rather than optimized text-image matching. While this choice reflects realistic vector database deployment scenarios, the Amazon Bedrock API offers a dedicated `IMAGE_RETRIEVAL` purpose that may yield stronger performance on text-to-image tasks; we leave this comparison to future work.

Precomputation - the practical case for embedding models. Embedding models are highly suit-

System A	System B	Recall@1 - McNemar p	Sig.	MRR - Wilcoxon p	Sig.
Gemini Embedding 2	GPT-4.1	0.2699	No	0.1574	No
Gemini Embedding 2	Claude Sonnet 4.6	0.5569	No	0.0995	No
GPT-4.1	Claude Sonnet 4.6	0.5854	No	0.9354	No
Gemini Embedding 2	Amazon Nova 2	<0.0001	**Yes**	<0.0001	**Yes**
GPT-4.1	Amazon Nova 2	<0.0001	**Yes**	<0.0001	**Yes**
Claude Sonnet 4.6	Amazon Nova 2	<0.0001	**Yes**	<0.0001	**Yes**

Table 1: Pairwise statistical significance tests. McNemar's test is applied to paired binary R@1 outcomes; Wilcoxon signed-rank test is applied to paired MRR scores. The upper block shows non-significant pairs; the lower block shows pairs involving Amazon Nova 2, all significant at $p < 0.0001$.

	Precomputation Time			
System	**Query emb.**	**Image emb.**	**Total Time**	**Ranking-only Time**
Gemini Embedding 2	∼2,000 s	∼12,500 s	∼14,500 s	**0.66 s** (with precomp.)
Amazon Nova 2	∼1,800 s	∼10,000 s	∼11,800 s	**1.16 s** (with precomp.)
GPT-4.1	N/A (end-to-end per-query API calls)			6,100 s (no precomp.)
Claude Sonnet 4.6	N/A (end-to-end per-query API calls)			9,415 s (no precomp.)

Table 2: Execution time estimates for 1,000 queries with 25 candidates. *Precomputation Time*: GE2: ∼2.0 s/query, ∼2.5 s/image; Amazon Nova 2: ∼1.8 s/query, ∼2.0 s/image; the image embedding time is aggregated over 5,000 unique candidate images. *With precomputation*: embeddings are computed once and reused for ranking.

Figure 4: Gemini Embedding 2 model is correct; others incorrect.

able for ranking large data with lower latency owing to precomputation, compared to using LLMs as visual rankers which are limited to a smaller candidate set. Without precomputation, embedding model total time is broadly comparable to LLM end-to-end time (depending on candidate pool overlap). Table 2 shows that once precomputation is complete, online ranking latency (for 1000 queries and 25 candidates) drops to under 2 seconds for both Gemini Embedding 2 and Amazon Nova 2, compared to 6,100 seconds and 9,415 seconds for GPT-4.1 and Claude Sonnet 4.6, respectively. This speedup is four orders of magnitude faster. For applications requiring sub-second retrieval over a large but fixed set of text and/or images, embedding models are the better choice (especially GE2 because of high recall scores). One critical distinction is that embedding model costs are incurred *once*, when creating the embeddings, and amortized across all subsequent queries. On the other hand, LLM costs scale linearly with every query. LLM-based rankers may be preferable when the image gallery is frequently updated or small, rendering precomputation impractical or its amortized benefit negligible.

5 Conclusion

We have presented a comparison of natively multimodal embedding models (Gemini Embedding 2 and Amazon Nova 2), and LLMs (GPT-4.1 and Claude-Sonnet-4.6) for text-to-image retrieval, using hard-negative candidates from Flickr30k. Gemini Embedding 2, GPT-4.1, and Claude Sonnet 4.6 achieve statistically indistinguishable retrieval accuracy on all three metrics - Recall@1, Recall@3, and MRR. The practical implication favors multimodal embedding models for low-latency applications. Once embeddings are precomputed, Gemini Embedding 2 and Amazon Nova 2 rank 1,000 queries in under two seconds, compared to 6,100 and 9,415 seconds for GPT-4.1 and Claude Sonnet 4.6, respectively. Extending this comparison to additional datasets and to video and audio modalities are natural directions for future work.

Limitations

Caption–image correspondence. Flickr30k captions describe images at varying levels of specificity, and a caption from one image may plausibly describe another image in the dataset, particularly among hard negatives selected for semantic sim-

ilarity. This creates ambiguous ground truth for some queries. We argue this limitation affects all evaluated systems equally, as none is given privileged information about which image served as the caption's source.

Hard negatives defined in caption space. Candidate sets are hard with respect to caption-level semantics but the "hardness" may change in image embedding space. A distractor that is semantically close in caption space may or may not be visually confusable in the embedding or reasoning space of a given model. This is a deliberate design choice, as discussed in Section 3. Another limitation of selecting hard negatives based on caption similarity is that it may inherently favor embedding-based methods optimized for dense vector representations.

Different inference conditions. The two paradigms operate under fundamentally different inference conditions: LLMs receive all 25 candidate images simultaneously and may exploit cross-candidate comparisons before producing a ranking, while embedding models score each image independently without access to other candidates.

Proprietary model access. All four evaluated systems are accessible only through paid APIs, and model weights may be updated by providers between our experiments and any replication attempt.

Candidate set size and generalizability. With 25-image candidate sets, our evaluation measures ranking performance rather than full-gallery retrieval. Conclusions about relative system performance may not generalize to large-scale settings.

LLM Ranking Prompt and Replicability. Changes to LLM prompt for GPT-4.1, and Claude-Sonnet-4.6 may affect the ranking. For reproducibility, we have reported the prompt we created, in Appendix C. Additionally, candidates could be presented to LLMs in smaller batches rather than all 25 simultaneously, which may mitigate position bias.

References

Amazon. 2025. Amazon nova multimodal embeddings: Technical report and model card. Technical report, Amazon Web Services.

Anthropic. 2024. Claude: A family of large language models. Available at https://www.anthropic.com.

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. In *Advances in Neural Information Processing Systems*, volume 36.

Bradley Efron and Robert J. Tibshirani. 1993. *An Introduction to the Bootstrap*. Chapman and Hall/CRC.

Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. VSE++: Improving visual-semantic embeddings with hard negatives. In *British Machine Vision Conference (BMVC)*.

Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2023. EVA-CLIP: Improved training techniques for CLIP at scale. *arXiv preprint arXiv:2303.15389*.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. ImageBind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190.

Google DeepMind. 2026. Gemini Embedding 2: Our first natively multimodal embedding model. Available at https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-embedding-2/. Accessed March 2026.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-hsuan Sung, Zhen Li, and Tom Duerig. 2021. ALIGN: Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 4904–4916. PMLR.

Hongwei Jiang, Tengyu Lian, and 1 others. 2024. MM-Embed: Universal multimodal retrieval with multimodal LLMs. *Preprint*, arXiv:2411.02571.

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz Castillo, Semih Yavuz, and Wenhu Chen. 2025. VLM2Vec: Training vision-language models for massive multimodal embedding tasks. In *International Conference on Learning Representations*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, pages 19730–19742. PMLR.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755. Springer.

Xueguang Ma, Xinyu Zhang, Ronak Pradeep, and Jimmy Lin. 2023. Zero-shot listwise document reranking with a large language model. *arXiv preprint arXiv:2305.02156*.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

OpenAI. 2025. GPT-4.1 system card. Technical report, OpenAI.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 3982–3992.

Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. 2021. Contrastive learning with hard negative samples. In *International Conference on Learning Representations*.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is ChatGPT good at search? investigating large language models as re-ranking agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14918–14937.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*.

Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Weizhi Zhu and 1 others. 2024. MLLM is a strong reranker: Advancing multimodal retrieval-augmented generation via knowledge-enhanced reranking and noise-injected training. *Preprint*, arXiv:2407.21439.

A Details of Evaluated Systems

System 1: Gemini Embedding 2 (Gemini Embedding 2). We embed each query caption using the `gemini-embedding-2-preview` model with `task_type=RETRIEVAL_QUERY` and embed each candidate image with `task_type=RETRIEVAL_DOCUMENT`, using 3,072-dimensional output. We compute cosine similarity between the caption embedding and each of the 25 candidate image embeddings, and rank the candidates from highest to lowest similarity.

System 2: Amazon Nova 2 - Multimodal Embedding 2. We embed each query caption using the `amazon.nova-2-multimodal-embeddings-v1:0` model with `embedding_purpose=GENERIC_INDEX` and also embed each candidate image with `embedding_purpose=GENERIC_INDEX`, using 3,072-dimensional output. We compute cosine similarity between the caption embedding and each of the 25 candidate image embeddings and rank the candidates from highest to lowest similarity.

System 3: GPT-4.1 We send the query caption and all 25 candidate images in a single prompt (zero-shot) to `gpt-4.1-2025-04-14` (latest snapshot at time of experiment). Images are presented in randomized order per query to control for position bias (Wang et al., 2023), numbered 1 through 25. The prompt instructs the model to return a ranking of all 25 images by their relevance to the caption, formatted as a list of image numbers. Randomization is applied independently per query, and the same random seed is used across both LLM conditions for reproducibility.

System 4: Claude Sonnet 4.6 We replicate the same procedure as GPT-4.1 for the `claude-sonnet-4-6 model` (Anthropic, 2024).

B Metrics with Confidence Intervals

Table 3 reports the complete metric results for all four systems with 95% bootstrap confidence intervals (1,000 resamples) and total ranking execution time for 1,000 queries.

C LLM Ranking Prompt

Here's the LLM Ranking prompt used:

```
"system": "You are a precise image retrieval system. Given a text query and a set of candidate images, rank ALL images from most to least relevant. Relevance means how well the image matches the specific subjects, actions, setting, and details described in the query. You must return a JSON array containing every image ID exactly once. Output ONLY the JSON
```

System	R@1	95% CI	R@3	95% CI	MRR	95% CI
GE2	0.804	[0.780, 0.829]	0.921	[0.905, 0.937]	0.871	[0.854, 0.888]
Nova2	0.673	[0.644, 0.703]	0.844	[0.821, 0.867]	0.774	[0.753, 0.796]
GPT-4.1	0.789	[0.763, 0.816]	0.915	[0.898, 0.932]	0.859	[0.841, 0.877]
Claude Sonnet 4.6	0.796	[0.772, 0.823]	0.903	[0.886, 0.921]	0.857	[0.839, 0.876]

Table 3: Full results for all systems on 1,000 queries with 25 hard-negative candidates per query. Bootstrap CIs are 95% intervals computed over 1,000 resamples. Random baseline values are theoretical (1/25 for R@1, 3/25 for R@3, harmonic mean for MRR).

`array — no explanation, no commentary, nothing else."`

`"intro_template": "Text query: query_caption. You are given n_candidates candidate images, each labeled with an image ID shown before it. For each image, consider: (1) Does it show the same subjects or people described in the query? (2) Does it show the same action or activity? (3) Does it match the setting, location, or scene? (4) Does it contain the specific objects or details mentioned? Images:"`

`"outro": "Now return a JSON array of ALL n_candidates image IDs ranked from most relevant (index 0) to least relevant (last index). Every ID must appear exactly once. If two images seem equally relevant, use your best judgment to order them — do not omit any ID. Provide JSON array only:",`

D More Hard Negative Dataset Examples

Figures 5–7 illustrate representative hard negative candidate sets. Each figure shows the query caption, the correct image (green border), and the top-k most semantically similar images as selected by `all-mpnet-base-v2` caption similarity. These examples demonstrate that the hard negatives are visually and semantically confusable with the correct image, validating the difficulty of the evaluation task.

E Contingency Tables

Table 4 shows the R@1 contingency table for GE2 versus GPT-4.1. Of the 1,000 queries, 716 are answered correctly by both systems, while 88 are answered correctly by GE2 alone and 73 by GPT-4.1 alone.

Figure 5: Hard negative example 1. The query caption describes a specific scene; candidate images share overlapping visual and semantic elements that make the retrieval task non-trivial.

Figure 6: Hard negative example 2.

	GPT-4.1 correct	GPT-4.1 wrong
GE2 correct	716	88
GE2 wrong	73	123

Table 4: R@1 contingency table for Gemini Embedding 2 vs. GPT-4.1 (1,000 queries). Off-diagonal cells represent qualitatively distinct failure modes analysed in Section 4. Contingency tables for all other system pairs are in Appendix E.

Tables 5–9 report R@1 contingency tables for the remaining system pairs not shown in the main text. Off-diagonal cells indicate queries where the two systems disagree, forming the basis for qualitative error analysis.

	Claude correct	Claude wrong
GE2 correct	729	75
GE2 wrong	67	129

Table 5: R@1 contingency table: GE2 vs. Claude Sonnet 4.6 (1,000 queries).

Figure 7: Hard negative example 3.

	Claude correct	Claude wrong
GPT-4.1 correct	732	57
GPT-4.1 wrong	64	147

Table 6: R@1 contingency table: GPT-4.1 vs. Claude Sonnet 4.6 (1,000 queries).

	Nova2 correct	Nova2 wrong
GE2 correct	634	170
GE2 wrong	39	157

Table 7: R@1 contingency table: GE2 vs. Nova2 (1,000 queries).

	GPT-4.1 correct	GPT-4.1 wrong
Nova2 correct	617	56
Nova2 wrong	172	155

Table 8: R@1 contingency table: GPT-4.1 vs. Nova2 (1,000 queries).

	Claude correct	Claude wrong
Nova2 correct	624	49
Nova2 wrong	172	155

Table 9: R@1 contingency table: Claude Sonnet 4.6 vs. Nova2 (1,000 queries).

F Qualitative Observation

Each figure shows the query caption, the correct image, and the top-ranked image returned by each system. Green text indicate a correct rank-1 prediction; red text indicate an incorrect rank-1 prediction. Figure 8 is another example where GE2 selected the correct image despite it being dark. Figure 9 shows how close the systems are to each other in terms of retrieval for certain queries. Even though GE2 was technically correct, the other systems also picked "reasonably correct" images (based on human vision). This illustrates that the differences in retrieval capability of each system may be lesser than what the quantitative metrics convey.

Interestingly, Figure 10 shows an example when Nova 2 selected the correct image while all others were incorrect. This indicates that Nova 2 may be better for certain queries, which can be explored further. Figure 11 and 12 show where only the Multimodal Embedding models were correct and only LLMs were correct, respectively. The following figures provide more examples on the performance of evaluated systems.

F.1 GE2 correct, all LLMs wrong

Figure 13 shows queries where GE2 ranked the correct image first while GPT-4.1 and Claude Sonnet 4.6 both failed. These cases tend to involve fine-grained visual similarity where the correct image shares compositional structure or low-level visual properties with the distractors that GE2's embedding captures but LLMs miss when reasoning jointly over 25 candidates.

F.2 Nova2 correct, all other systems wrong

Figure 14 shows the rare queries where Nova2 ranked correctly while GE2, GPT-4.1, and Claude Sonnet 4.6 all failed. Given Nova2's overall lower performance, these cases are uncommon but analytically interesting as they may reflect query types where GENERIC_INDEX embeddings capture complementary signal.

F.3 Both LLMs correct, both embedding models wrong

Figure 15 shows queries where both GPT-4.1 and Claude Sonnet ranked correctly while both GE2 and Nova2 failed. These cases illustrate the advantage of cross-candidate joint reasoning: the LLM can compare all 25 images simultaneously and apply semantic or relational understanding that independent embedding scoring cannot replicate.

F.4 Both embedding models correct, both LLMs wrong

Figure 16 shows queries where both GE2 and Nova2 ranked correctly while both LLMs failed. These cases are particularly informative as they suggest that dense cross-modal embedding captures visual correspondence that joint visual reasoning consistently misses, even when the pattern holds across model families.

Figure 8: Another example where Gemini Embedding 2 model is correct; others incorrect.

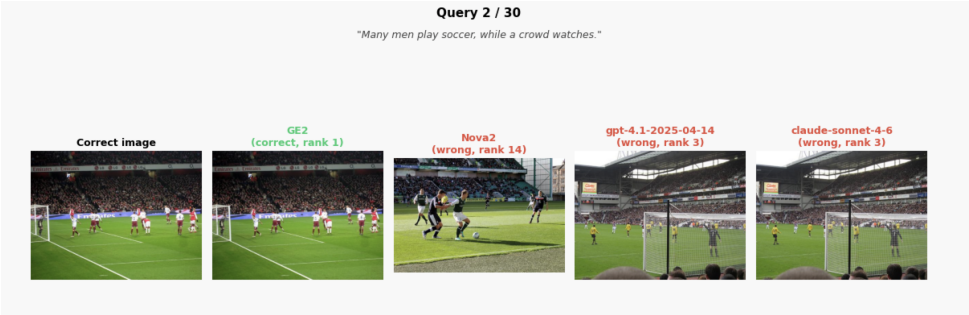


Figure 9: An example where all models' retrieval reasonably close to the query.

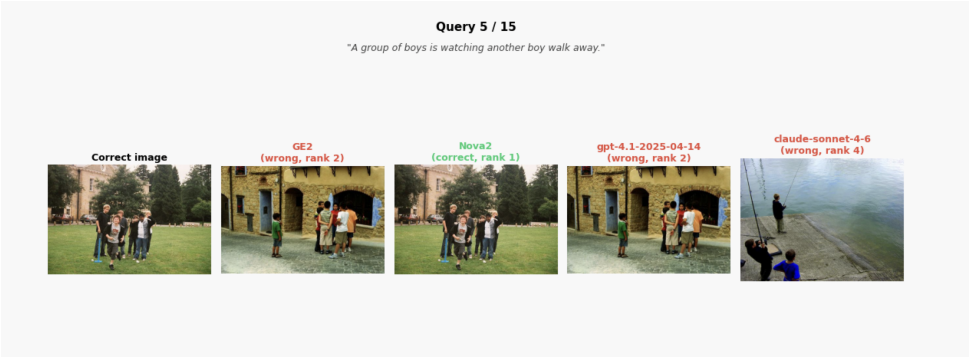


Figure 10: Amazon Nova 2 results are correct; Other models' results are incorrect.

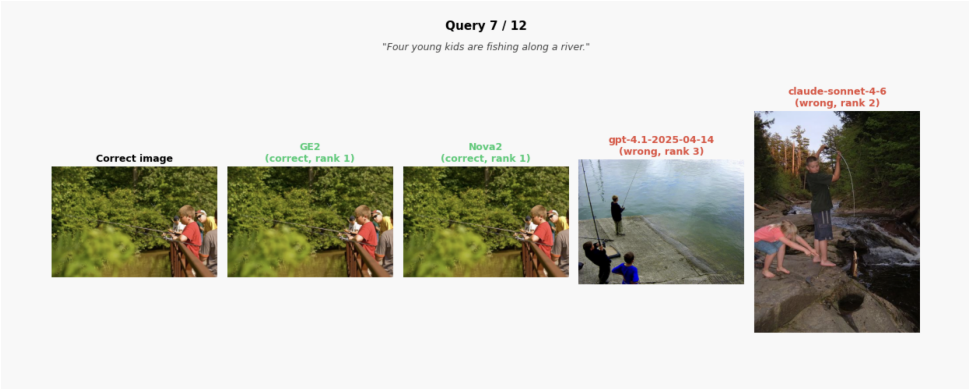


Figure 11: Multimodal embedding models are correct; LLM Rankers incorrect.

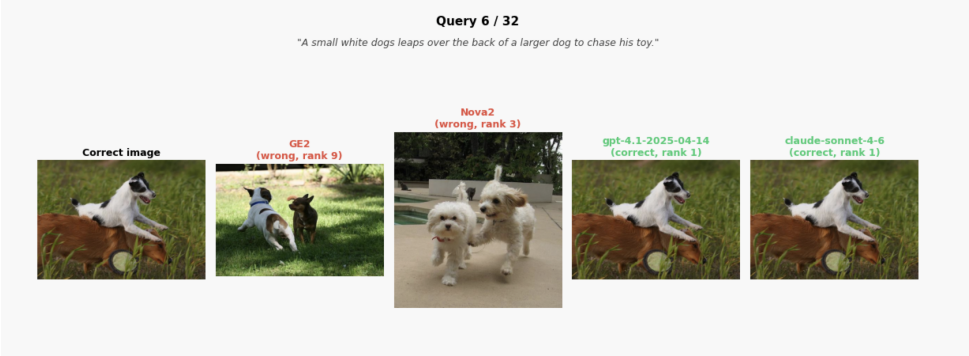
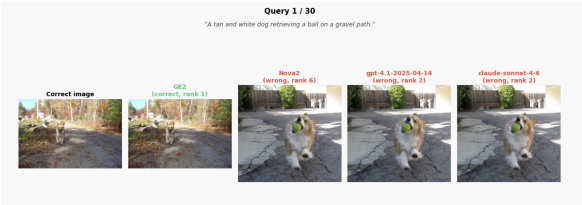
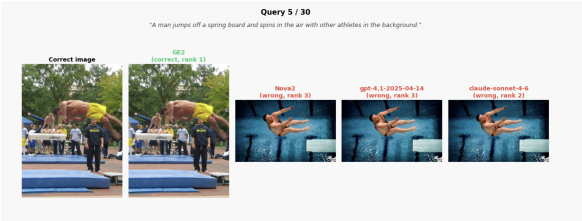


Figure 12: LLM models' results are correct; Embedding models' results are incorrect.



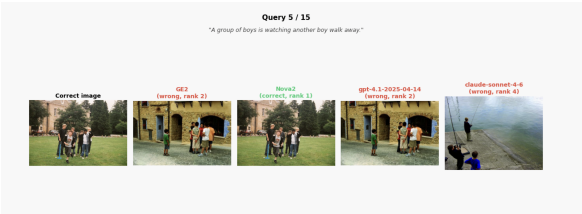
(a) Example 1



(b) Example 2

Figure 13: Queries where GE2 is correct and both LLMs (GPT-4.1 and Claude Sonnet) are wrong on R@1. Green = correct rank-1, red = incorrect rank-1.

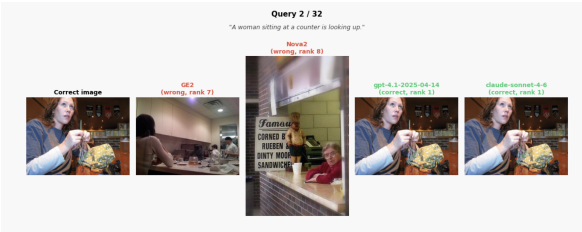
(a) Example 1



(b) Example 2

Figure 14: Queries where Nova2 is correct and all other systems are wrong on R@1.

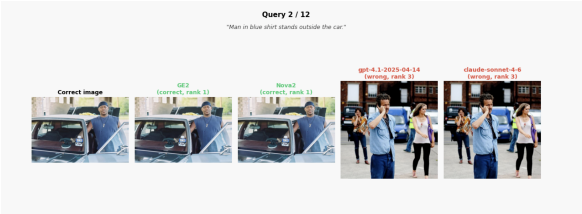
(a) Example 1



(b) Example 2

Figure 15: Queries where both LLMs (GPT-4.1 and Claude Sonnet 4.6) are correct and both embedding models (GE2 and Nova2) are wrong on R@1.

(a) Example 1



(b) Example 2

Figure 16: Queries where both embedding models (GE2 and Nova2) are correct and both LLMs (GPT-4.1 and Claude Sonnet 4.6) are wrong on R@1.