MMCOMET: A LARGE-SCALE MULTIMODAL COMMONSENSE KNOWLEDGE GRAPH FOR VISUAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Commonsense knowledge is widely used to support multimodal reasoning and generation, yet existing resources are predominantly text-only and lack grounding in visual context. This limitation often leads to the retrieval of semantically related but visually misaligned knowledge, and can degrade downstream performance. We introduce **MMCOMET**, a large-scale multimodal commonsense knowledge graph that integrates physical, social, and eventive knowledge with visual grounding. MMCOMET extends ATOMIC2020 by associating each commonsense triple with representative images through a hybrid alignment pipeline, resulting in over 900K multimodal tuples spanning 19 relation types. We further establish a unified evaluation framework across diverse vision-language tasks, including visual commonsense reasoning, visual question answering, visual storytelling, image editing, and image captioning. Our analysis shows that multimodal grounding provides more contextually aligned and reliable knowledge compared to both text-only knowledge graphs and LLM-based knowledge generation, particularly in reasoning-intensive scenarios. At the same time, we observe that the effectiveness of external knowledge is task-dependent, and naive knowledge integration can introduce noise in perception-dominated settings. To mitigate this, we propose an adaptive knowledge router that dynamically selects optimal paths or bypass knowledge in harmful settings. MMCOMET provides a scalable multimodal commonsense resource and a benchmark for studying the role of grounded knowledge in multimodal systems.[1]

1 INTRODUCTION

Commonsense knowledge graphs (KGs) encode everyday knowledge as relational triples $\langle \texttt{head}, \texttt{relation}, \texttt{tail} \rangle$ and have become important external resources for retrieval-augmented reasoning and generation (Guu et al., 2020). In multimodal settings, such knowledge provides auxiliary context for tasks including visual commonsense reasoning (VCR) (Zellers et al., 2019), visual question answering (VQA) (Gardères et al., 2020), visual storytelling (VST) (Wang et al., 2024), image editing (IE) (Ju et al., 2024), and image captioning (IC) (Zhou et al., 2019). In these pipelines, the quality and relevance of retrieved knowledge directly affect downstream outputs (Razniewski et al., 2024). However, most existing commonsense KGs remain text-only and therefore cannot directly support visually grounded knowledge retrieval.

Recent multimodal knowledge graphs (MMKGs) incorporate images into large-scale encyclopedic resources (Ferrada et al., 2017; Alberts et al., 2021). Although effective for factual and entity-centric queries, they primarily represent encyclopedic relations, such as *located-in* and *capital-of*, and provide limited coverage of everyday commonsense involving social interactions, event dynamics, and human intentions. Large-scale commonsense KGs such as ATOMIC2020 (Hwang et al., 2021) capture rich social, physical, and eventive knowledge, but remain entirely textual. Existing resources thus do not jointly represent inferential commonsense and its visual grounding. Another approach is to generate commonsense knowledge directly using large language models (LLMs). While flexible, LLM-generated knowledge may be ungrounded, irrelevant, or inconsistent with the visual input. Similarly, text-only KGs can retrieve knowledge that is semantically related to a query but visually misaligned with the depicted situation. As illustrated in Figure 1, systems without knowledge or

[1] Code and data are available at `https://anonymous.4open.science/r/MMCOMET_CODE-4595`

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

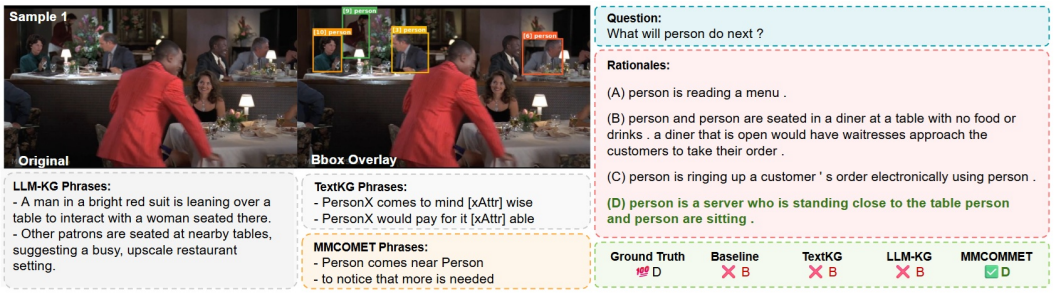


Figure 1: An example of visual commonsense reasoning featuring results from no knowledge injected (baseline), text-based knowledge graph injection (TextKG), LLM-generated knowledge injection (LLM-KG), and MMCOMET knowledge injection.

with text-only knowledge may therefore produce generic or weakly grounded reasoning. Multimodal commonsense can provide more contextually aligned inference by jointly capturing (i) physical knowledge about objects and environments, (ii) eventive knowledge about actions and dynamics, and (iii) social knowledge about roles, intentions, and interactions.

To address this gap, we introduce **MMCOMET**, a large-scale multimodal commonsense KG that integrates physical, social, and eventive knowledge with visual grounding. MMCOMET extends ATOMIC2020 by associating commonsense phrases with representative images through a hybrid pipeline that combines embedding-based retrieval with web-scale search. The resulting resource contains approximately 1M multimodal commonsense tuples spanning 19 relation types. Beyond constructing the resource, we establish a unified evaluation framework for studying multimodal commonsense across VCR, VQA, VST, IE, and IC. We compare MMCOMET with text-only KG and LLM-generated knowledge. Our experiments show that visually grounded commonsense provides more contextually aligned and reliable knowledge, particularly for reasoning-intensive tasks. At the same time, the benefits of external knowledge vary across tasks, indicating that effective knowledge integration requires adaptive knowledge selection rather than naive integration.

Our contributions are summarized as follows: **1)** We introduce MMCOMET, a large-scale multimodal commonsense knowledge graph that integrates physical, social, and eventive knowledge with visual grounding, comprising approximately 1M multimodal tuples. **2)** We propose a scalable hybrid alignment pipeline that associates commonsense phrases with representative images using embedding-based retrieval and web search. **3)** We provide comprehensive dataset statistics and human evaluation validating the quality and relevance of the resulting image–text alignments. **4)** We establish a unified evaluation framework across five vision-language tasks and analyze multimodal commonsense against text-only KG retrieval and LLM-generated knowledge. **5)** We propose an adaptive knowledge router that dynamically selects optimal knowledge paths and prevents harmful interventions when no route provides positive utility. MMCOMET serves as a reusable resource for multimodal knowledge retrieval, reasoning, and generation, as well as a benchmark for studying grounded commonsense in vision-language systems.

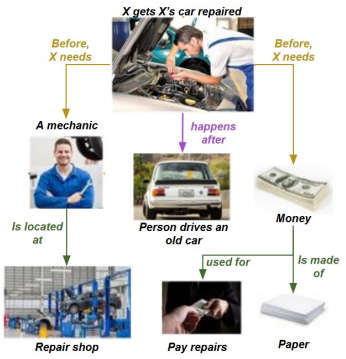
2 Related Work

Text-only Commonsense Knowledge Graphs Commonsense knowledge graphs (CSKGs) encode everyday knowledge beyond encyclopedic facts. Existing resources have been constructed through crowdsourcing (Liu & Singh, 2004; Speer et al., 2017), web-scale extraction (Tandon et al., 2014; Nguyen et al., 2023), and neural scoring (Zhang et al., 2020). Among them, ATOMIC and ATOMIC2020 (Sap et al., 2019; Hwang et al., 2021) provide large-scale inferential knowledge about social interactions, events, and physical concepts, supporting tasks such as dialogue and narrative reasoning (Cai et al., 2023; Wang et al., 2024). However, these resources remain text-only and cannot assess whether retrieved commonsense knowledge is visually aligned with a given scene.

Multimodal Knowledge Graphs Multimodal knowledge graphs (MMKGs) associate structured knowledge with images or other modalities. Most existing MMKGs ground encyclopedic entities and factual relations from resources such as DBpedia, Freebase, and Wikidata (Ferrada et al., 2017; Liu et al., 2019; Wang et al., 2020; Wu et al., 2023; Gong et al., 2024). A smaller set targets commonsense knowledge: MMEKG (Ma et al., 2022) focuses on multimodal events, TIVA-KG (Wang et al., 2023) emphasises physical commonsense, and VTKG-C (Lee et al., 2023) grounds visually expressible entities and triplets. As summarised in Table 1, these resources do not jointly cover

Table 1: Comparison of existing MMKGs with MMCOMET (left) and a small subset of MMCOMET (right). MMCOMET uniquely covers domains (social, physical, eventive) with nearly 1M tuples (Ency: Encyclopedic; Geo: Geographic; Multi-L: Multilingual; Edu: Education; CS: Commonsense; So: Social; Ph: Physical; Ev: Eventive; C: Concept; E: Entity; R: Relation; T: Tuples; WSE: Web Search Engine; WP: Wikipedia; WMC: Wikimedia Commons; OID: open-source image datasets; ImgNet: ImageNet.)

MMKG	Domain	Size	MultiM	Image
IMGpedia (Ferrada et al., 2017)	Ency	442M	E, R	WMC
ImageGraph (Oñoro-Rubio et al., 2019)	Ency	560K	E	WSE
MMKG (Liu et al., 2019)	Ency	814K	E	WSE
MMpedia (Wu et al., 2023)	Ency	19M	E	WSE, WP
AspectMMKG (Zhang et al., 2023)	Ency	645K	E	WSE, WP
UKnow (Gong et al., 2024)	Ency	3.6M	E, T	WSE, WP
Richpedia (Wang et al., 2020)	Ency, Geo	172M	E, R	WSE, WP
VisualSem (Alberts et al., 2021)	Ency, Multi-L	1.5M	E	WP, ImgNet
SciMKG (Lu et al., 2026)	Edu	403K	C, R, T	WP, OID
MMEKG (Ma et al., 2022)	CS (Ev)	934M	E, T	imSitu
TIVA-KG (Wang et al., 2023)	CS (Ph)	1.4M	E, T	WSE
VTKG-C (Lee et al., 2023)	CS (Ph)	111K	E, R, T	OID
MMCOMET (*ours*)	CS (So, Ph, Ev)	989K	E, R	WSE, OID



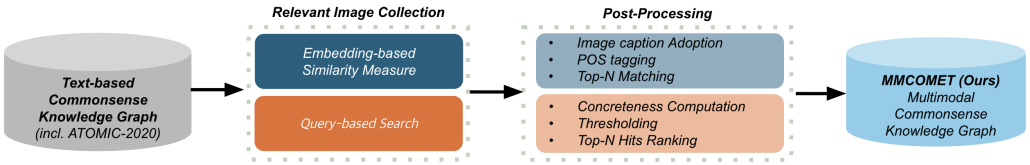


Figure 2: The MMCOMET construction pipeline for grounding a text-based commonsense knowledge graph with representative visual evidence.

social, physical, and eventive commonsense at scale. MMCOMET addresses this gap by visually grounding nearly 1M ATOMIC2020 triples, providing a unified resource for multimodal commonsense retrieval, reasoning, and generation.

3 MMCOMET

3.1 Base Commonsense Topology

MMCOMET is constructed by extending the textual commonsense graph ATOMIC2020 (Hwang et al., 2021) with a visually grounded layer. For each triple $\langle \texttt{head}, \texttt{relation}, \texttt{tail} \rangle$, we associate representative images for the head and tail phrases through a hybrid retrieval pipeline. Table 1 shows a subset of MMCOMET (right). Figure 2 presents the overall construction workflow. We adopt ATOMIC2020 as the textual backbone, which contains $\sim$1.33M inferential commonsense triples spanning social, physical, and eventive relations. After filtering incomplete tuples and non-visualizable relation types, including the `isFilledBy`, which contains blank placeholders, MMCOMET retains 989K multimodal tuples spanning 19 relation categories. MMCOMET does not aim to alter the relational structure, but to *ground* each textual node with representative visual evidence, enabling multimodal reasoning.

3.2 Hybrid Visual Alignment Pipeline

Associating images with commonsense phrases presents two challenges: (i) physical concepts are often visually concrete and (ii) social and eventive concepts are frequently abstract. To address this, we design a two-stage retrieval strategy:

Embedding-based Similarity Matching For visually concrete phrases (e.g., `ObjectUse`), we adopt embedding-based retrieval. Head and tail phrases are encoded using CLIP (ViT-B/32) (Radford et al., 2021) text encoders and matched via cosine similarity against pre-encoded image embeddings from open-source captioning and storytelling datasets Conceptual Captions (Sharma et al., 2018), COCO (Lin et al., 2014), Flickr30K (Young et al., 2014), and the Visual Storytelling dataset (Huang et al., 2016). To ensure scalability, we introduce a noun-indexed filtering mechanism. Image captions are POS-tagged using NLTK (Bird et al., 2009) to extract noun tokens, which are organized into an inverted index, mapping nouns to candidate images. Similarity is first computed

Table 2: (Left) Human evaluation results showing the percentage of commonsense phrases whose retrieved images were judged as full (FM) or partial matches (PM). (Right) Average concreteness scores and sample sizes.

	Web (WSE)		Open-source Data (OID)		Both		Relation Type Stats.	
Physical	**FM+PM**	**FM**	**FM+PM**	**FM**	**FM+PM**	**FM**	**Avg. Concreteness**	**Size**
ObjectUse	100	73	96	76	98	74.5	3.90	165,590
AtLocation	100	92.9	100	100	100	96.4	4.15	20,221
MadeUpOf	100	100	89.3	89.3	94.1	94.1	4.08	3,345
HasProperty	92.6	92.6	88.5	84.6	90.6	88.7	3.80	5,617
CapableOf	95.8	87.5	95.8	70.8	95.8	79.2	3.96	7,968
Desires	100	95.7	100	72	100	83.3	3.88	2,737
NotDesires	100	92.6	88	72	94.2	82.7	3.91	2,838
Event	**FM+PM**	**FM**	**FM+PM**	**FM**	**FM+PM**	**FM**	**Avg. Concreteness**	**Size**
isAfter	91.3	69.6	77.3	45.5	84.4	57.8	3.94	22,453
HasSubEvent	96.3	92.6	92.9	71.4	94.5	81.8	3.58	12,845
isBefore	87.5	62.5	88.9	63	88.2	62.7	3.91	23,208
HinderedBy	86.7	73.3	76.2	42.9	80.6	55.6	3.80	106,647
Causes	96.2	96.2	87.5	79.2	92	88	3.29	376
Reason	100	92	83.3	75	91.8	83.7	3.55	334
Social	**FM+PM**	**FM**	**FM+PM**	**FM**	**FM+PM**	**FM**	**Avg. Concreteness**	**Size**
Need	90.5	81	85.7	71.4	88.1	76.2	3.65	89,391
Attr	92	80	89.5	78.9	90.9	79.5	2.98	115,815
Effect	95	90	61.9	42.9	78	65.9	3.53	113,494
React	90.9	86.4	61.1	55.6	77.5	72.5	2.95	94,228
Want	80.8	73.1	86.4	50	83.3	62.5	3.58	152,989
Intent	95	80	88.2	58.9	91.9	70.3	3.40	49,312
All Relations	91.5	79.0	82.4	61.2	86.9	70.1	3.55	989,408

at the noun level to identify a reduced candidate subset, after which fine-grained phrase-to-image similarity is applied.

Concreteness-aware Web Retrieval For abstract social and eventive phrases (e.g., `React`, `Intent`, `Want`), caption-based similarity matching often fails to produce adequate visual grounding. We employ a concreteness-aware routing strategy. We compute phrase-level concreteness scores using the 40K Concreteness Ratings (Brysbaert et al., 2014). Phrases with an average concreteness score below 4.0, a threshold aligned with the empirical distribution of social and eventive relations, are routed to web-based image retrieval. Structured web queries are issued with entity normalization and text-filtering heuristics to reduce noise. Approximately 72% of unique phrases are handled through this abstract-concept routing process.

3.3 Post-processing and Image Selection

All candidate images (from similarity matching and web retrieval) are re-ranked using CLIP similarity with the original phrase. Low-alignment samples are filtered using a threshold of 0.15, and the top 10–15 images per phrase are retained. The final MMCOMET graph consists of 989K multimodal commonsense triples, 578K unique textual nodes, and $\sim$3.9M associated images in social, physical, and eventive relations. This hybrid pipeline enables scalable multimodal grounding while maintaining high alignment quality, forming a retrieval-ready multimodal commonsense resource. We provide extensive details, statistics, examples, retrieval analysis, and error analysis of the MMCOMET graph construction in the Appendix. D.

4 Intrinsic Evaluation of MMCOMET: Multimodal Alignment

Human Evaluation We conduct a human evaluation to assess the semantic alignment between commonsense phrases and their aligned images in MMCOMET. For each of the 19 relations, 60 phrases are randomly sampled, resulting in 1140 phrases divided among three annotators (Appendix E). To balance retrieval strategies, 50% of the samples are drawn from phrases that required web search. For each head/tail phrase, annotators are shown the top 7 images, yielding 7980 phrase-image pairs. Each image is then labeled as either 1) *Full Match (FM)*, where all key concepts expressed in the phrase are visually present; 2) *Partial Match (PM)*, where the image is thematically relevant but missing some elements; or 3) *No Match (NM)*, where the image does not correspond to the phrase. Additionally, annotators rate concreteness as *Low*, *Medium*, or *High*, corresponding to the difficulty of visually grounding the phrase. In particular, *Low* denotes abstract or intangible items, while *High* are concrete, tangible objects. Examples for each semantic alignment label may be found in the Appendix. J.2.

Table 2 reports the percentage of phrases for which at least 5 out of the top 7 retrieved images were judged as either Full Match (FM) or Partial Match (PM). Across all relations and image sources,

MMCOMET achieves 86.9% FM+PM alignment and 70.1% FM alignment, indicating that the majority of commonsense phrases are grounded with visually relevant evidence. Although Full Match scores are naturally lower, partially matched images remain semantically relevant and preserve core contextual elements, making them suitable for downstream retrieval and reasoning. Inter-annotator agreement is strong, with intraclass correlation coefficients of 0.76 (FM+PM) and 0.82 (FM), indicating consistent annotation quality. Overall, the human evaluation confirms that MMCOMET provides high-quality multimodal grounding across diverse commonsense relations, supporting its reliability as a retrieval-oriented multimodal resource.

Concreteness Analysis by Relation Type We analyze phrase-level concreteness scores across relation types to better understand grounding difficulty and the necessity of hybrid retrieval. The right sub-table of Table 2 reports average concreteness scores per relation. Physical relations exhibit consistently higher concreteness scores than social and eventive relations, with the exception of temporal relations such as `isBefore` and `isAfter`. Social-interaction relations show the lowest average concreteness, reflecting their reliance on abstract emotional and behavioral descriptions. For example, `React` frequently involves affective expressions, while `Attr` often consists of adjectival characterizations, both of which are inherently difficult to ground visually. We provide further analysis and examples in the Appendix. J.

5 Extrinsic Validation of MMCOMET: Downstream Task Evaluation

To assess the practical utility of MMCOMET as a reusable multimodal commonsense resource, we evaluate its impact on five vision-language tasks: Visual Commonsense Reasoning (VCR), Visual Question Answering (VQA), Visual Storytelling (VST), Image Editing (IE), and Image Captioning (IC)[2]. For each task, we provide extensive examples and qualitative analysis in the Appendix. I.

Experimental Setup For all tasks except Visual Storytelling, we compare three knowledge injection setups: 1) **Baseline** with no knowledge injected; 2) **TextKG**, which injects ATOMIC2020 text-based knowledge triplets into each task prompt; and 3) **MMCOMET**, which injects multimodal commonsense knowledge phrases incorporating physical, social, and eventive dimensions. We inject the top k knowledge triplets/phrases retrieved using L2 similarity between the image, the triplet/phrase, and the query for each task sample. We empirically set best k, where the performance is consistently strong across tasks. As the retrieval unit is identical across conditions, TextKG and MMCOMET receive an equal amount of injected knowledge. Exploring task-specific values of k remains a direction for future work. We evaluate Visual Storytelling by enriching the story graph with MMCOMET and compare it with prior work (Section 5.3). All experiments were conducted with single H100 GPU with 94GB GPU memory. Task-specific datasets and metrics are further discussed in each task section. Relevant model details, hyperparameters, and dataset statistics may be found in Appendix F. Prompt templates are provided in Appendix L.

5.1 Visual Commonsense Reasoning (VCR)

Visual Commonsense Reasoning is a vision-language task that requires models to select a correct answer given a question (Q→ A) or to select a rationale given a question-answer pair (QA→ R). In this work, we intentionally modify the task setup to require models to select a rationale from the question alone (Q→R), without access to the ground-truth answer, encouraging more implicit reasoning and preventing shortcut rationale selection. We evaluate the significance of multimodal physical, social, and eventive commonsense knowledge to baselines using the VCR benchmark dataset (Zellers et al.,

Table 3: VCR (Q→ R) accuracy with different settings. Δ values represents performance differences between TextKG (T) and MMCOMET (M) compared to the Baseline (B). MMCOMET-enhanced models outperform baselines in accuracy in both smaller and larger models.

Model	B	T	M	Δ_{T-B}	Δ_{M-B}
Qwen8B	65.3	62.7	**66.4**	-2.6	+1.1
Qwen32B	72.1	71.1	**72.4**	-1.0	+0.3
InternVL38B	73.8	74.1	**75.3**	+0.3	+1.5

[2]Across these tasks, we show that injecting MMCOMET's multimodal commonsense knowledge improves performance. However, this benefit is not uniform across all task types (See Sec. 6). We therefore do not claim that MMCOMET is broadly helpful for any vision-language task, but rather that it is most effective when a task requires commonsense-driven reasoning beyond direct visual perception.



Figure 3: Category breakdown from the VCR dataset with representative commonsense phrases and images from MMCOMET, highlighting the diversity of situational applications of our multimodal physical, social, and eventive commonsense knowledge.

2019), measure performance using Accuracy, and run experiments with Qwen3-VL variants and InternVL3 in zero-shot and few-shot settings. We report the best setups in Table 3.

Overall Results Across all model setups, TextKG injection degrades performance in Qwen3 models and yields marginal gains in InternVL3 (-2.7 to +0.3), suggesting that TextKG-retrieved knowledge, without visual grounding, tends to introduce noise that misleads some models. MMCOMET shows greater improvements over the baseline across all models (+0.3 to +1.5), suggesting that multimodal grounding, with physical, social, and eventive knowledge, reduces noise and leads to more stable, more reliable reasoning performance than text-only commonsense knowledge. We further observe that performance gains vary between model architectures and size. For Qwen3, despite having a lower baseline performance, the 8B variant shows greater gains than its 32B counterpart. On the other hand, InternVL3-38B shows the best baseline performance along with the highest gains. These suggest that MMCOMET is more robust across different model architectures and sizes, while also indicating the need to calibrate the injected knowledge to the specific model setup.

Ambiguity Analysis To further investigate when multimodal commonsense knowledge is most beneficial, we analyze model performance based on ambiguity, defined as $Ambiguity = 1 - (p(top_1) - p(top_2))$, where $p(top_1)$ and $p(top_2)$ denote the probabilities assigned to the top-1 and -2 answer candidates by the baseline model, respectively. A higher ambiguity score indicates that the model assigns similar probabilities to multiple candidates, reflecting samples where the correct answer is less distinguishable, or inherently harder samples. We find that MMCOMET improves accuracy and reduces uncertainty, outperforming Baseline and TextKG across different categories described in Figure 7. Further, the performance gap between MMCOMET and TextKG grows as ambiguity increases. TextKG becomes detrimental as sample difficulty increases, suggesting that visual grounding is critical for harder samples. More analysis may be found in Appendix H.1.

Qualitative Analysis Figure 1 shows a representative VCR example comparing different external knowledge injected into a Q→ R prompt. Across all other setups, only MMCOMET commonsense knowledge achieves the correct rationale for the question "What will person do next?" The TextKG-retrieved knowledge fails to associate any commonsense knowledge relevant to the events depicted in the image, whereas LLM-generated knowledge focuses more on general descriptions of the scene. MMCOMET's social and event-based commonsense knowledge enables it to generate more event-based knowledge ("Person comes near Person" and "to notice that more is needed"), which leads the model to the correct rationale. More examples may be found in Appendix I.1.

Category Breakdown We further demonstrate the extrinsic diversity of commonsense knowledge in MMCOMET by retrieving the top MMCOMET images and phrases for different categories in the VCR dataset. Figure 3 illustrates that MMCOMET can visually ground representative images for phrases relevant to each category, such as "No one is giving away any emotion" under *Mental-feeling* and "talk about something else with person Y" under *Activity-doing*. This highlights the versatility of MMCOMET across different situational, social, and eventive applications and across both visual and textual modalities. Further breakdown analysis is in Appendix H.1.

5.2 Visual Question Answering (VQA)

Visual Question Answering is a multimodal task in which the primary goal is to answer a question based on a given image. It requires an understanding of visual components and grounding to the textual semantics of the question. External knowledge tends to improve this alignment.

We evaluate the significance of our multimodal commonsense knowledge to the VQA task using MLLM models Qwen3 and Gemma, the VQAv2 (Goyal et al., 2017) validation dataset, and VQA accuracy (Mañas et al., 2024), which is a metric robust to inter-human variability. Compared to VCR, VQAv2 consists of more open-ended, perceptual questions that require physical or eventive commonsense knowledge rather than purely complex social reasoning.

Table 4: Automated evaluation results and a qualitative sample for VQA. MMCOMET-enhanced models outperform baselines on accuracy across different models.

Model	Baseline	TextKG	MMCOMET
Qwen8B	68.60	68.41	**69.74**
gemma31B	64.72	65.55	**66.31**



Table 4 illustrates that injecting MMCOMET consistently improves performance over the baseline across both models, while TextKG yields either marginal gains or slight degradation. This gap suggests that visual grounding is critical even for open-ended factual questions, where TextKG may introduce contextually irrelevant information. The qualitative sample further illustrates the benefit of injecting the eventive knowledge "getting from one location to another" in answering the given questions. These results indicate that MMCOMET generalises beyond social and intentional reasoning, providing reliable commonsense augmentation across a broader range of question types.

5.3 Visual Storytelling (VST)

Unlike the previous tasks, Visual Storytelling (Huang et al., 2016) requires synthesising a sequence of images to generate a coherent, cohesive storyline, often going beyond factual descriptions and requiring interpretation of visual social cues. We evaluate MMCOMET using the VIST benchmark (Huang et al., 2016), following SCO-VIST's framework (Wang et al., 2024). SCO-VIST's textual story graph is encoded into a multimodal graph by querying MMCOMET to find a matching image for each textual node. The story is then decoded by feeding the storyline, MMCOMET and VIST images into LLaVA (Liu et al., 2023). We compare our setup with the best model from SCO-VIST and a modified setup using VILT (Kim et al., 2021) as decoder. We measure performance using Perplexity, UNION (U) (Guan & Huang, 2020), RoViST (Wang et al., 2022) set, SPICE (S) (Anderson et al., 2016), BLEURT (B) (Sellam et al., 2020), and MoverScore (M) (Zhao et al., 2019). Averaged RoVist (R) and composite R+S+B+M+U are reported as integrated measures. More details may be found in Appendix C.

Table 5: Automatic evaluation results of visual storytelling. MMCOMET-enhanced models outperform baselines in visual grounding, coherence, and overall story quality metrics

Model	RoViST(R)	RSBMU	StoryLength
SCO-VIST	77.7	263.5	52.3
VILT	80.2	259.2	50.1
LLaVa	82.2	263.9	51.9
+MMCOMET^{S}	<u>82.5</u>	<u>264.4</u>	50.8
+MMCOMET^{S+I}	**83.5**	**272.0**	53.4



Table 5 illustrates that incorporating multimodal commonsense knowledge improves visual grounding, coherence, non-redundancy and semantic similarity (RoViST). LLaVA+MMCOMET^{S+I} achieves the highest overall composite score (272.0) and RoViST (83.5%), outperforming non-augmented baselines. Qualitative examples show that MMCOMET-enhanced stories exhibit richer event descriptions, clearer social interactions, and more contextual details compared to baseline. More samples and results may be found in Appendix I.3 and G.

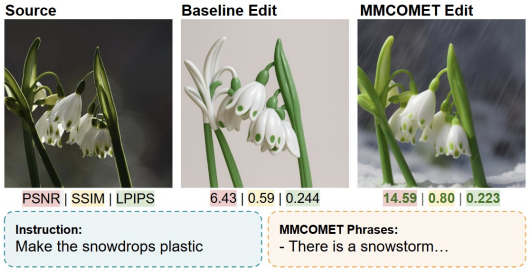
5.4 Image Editing (IE)

Image editing is the manipulation of an input image based on a given instruction. The task requires understanding the intent of the edit instruction and visually ground it to the input before generating a new image. To evaluate whether multimodal com-

monsense improves semantic consistency across various editing categories, we use the PIE_Bench dataset (Ju et al., 2024) and Qwen-Image-Edit (Wu et al., 2025). We measure edited image quality using PSNR, SSIM, and LPIPS. Table 6 illustrates that commonsense knowledge improves editing quality over Baseline. MMCOMET further improves gains from TextKG for SSIM (+0.0228) and LPIPS (-0.0134). We also provide qualitative examples (Appendix I.4) to assess the semantic consistency beyond image quality. While the baseline captures the intent of the instruction "Make the snowdrops plastic", it has changed the context of the source image and even removed the background. On the other hand, the MMCOMET edit retains the semantic attributes of the environment shown in the source image through the retrieved knowledge "There is a snowstorm...". Overall, MMCOMET not only improves the quality of edited images but also further establishes semantic consistency between the source and manipulated images.

Table 6: Automatic evaluation results and a qualitative sample for image editing. MMCOMET outperforms baselines in SSIM and LPIPS, achieves competitive PSNR, and demonstrates more faithful edits to the source and the instruction.

Metric	Baseline	TextKG	MMCOMET
PSNR ↑	25.7	**29.0025**	28.9692
SSIM ↑	0.86	0.9045	**0.9273**
LPIPS ↓	0.088	0.0728	**0.0594**



5.5 Image Captioning (IC)

Image captioning involves generating textual descriptions based on an input image. It requires identifying primary subject/s and understanding context of the image (Hossain et al., 2019), which could benefit from multimodal commonsense. We assess task performance using the Karpathy test split (Karpathy & Fei-Fei, 2017) of Flickr30k (Young et al., 2014), a widely used image captioning benchmark, across BLEU, CIDEr, and ROUGE-L metrics with fewshot prompting.

Table 7: Automatic evaluation results and a qualitative sample for Image Captioning. MMCOMET-enhanced models outperform other settings in all metrics.

Model	Metric	B	T	M
Qwen8B	BLEU-4	0.147	0.169	**0.170**
	CIDEr	0.232	0.346	**0.372**
	ROUGE-L	0.422	0.442	**0.446**
Qwen32B	BLEU-4	0.139	0.143	**0.146**
	CIDEr	0.151	0.147	**0.157**
	ROUGE-L	0.409	0.411	**0.413**
gemma31B	BLEU-4	0.255	0.257	**0.260**
	CIDEr	0.646	0.647	**0.650**
	ROUGE-L	0.526	0.527	**0.530**



Table 7 illustrates that knowledge-augmented setups consistently outperform baselines across all metrics. Further, incorporating MMCOMET's commonsense knowledge outperforms TextKG knowledge, highlighting the benefits of multimodal commonsense. Improvements in CIDEr indicate better alignment with the reference descriptions, while ROUGE-L gains suggest semantic richness. The qualitative sample demonstrates that MMCOMET introduces socially and eventively grounded interpretations (e.g. the intent of "finding a quiet place to read"), leading to more informative and context-aware captions.

5.6 Extrinsic Human Evaluation

Beyond automatic metrics, we conduct a human evaluation across four tasks (VCR, VQA, Image Editing, Image Captioning) to assess whether the retrieved commonsense knowledge is perceived as relevant and useful. Annotators rated 20 samples per task on a 4-point Likert scale for **(1) relevance of the retrieved knowledge to the image and input**, and **(2) its helpfulness for reasoning.** Average scores were 3.1 and 2.9, respectively, both above the neutral midpoint, and annotators perceived the retrieved knowledge as helpful even in cases where MMCOMET did not outperform the baseline on automatic metrics, suggesting a gap between automatic evaluation and human judgment. Full annotation details are provided in the Appendix E.2.

5.7 Effect of Knowledge Source

To further demonstrate the effectiveness of MMCOMET, we compare it against two alternative knowledge sources: LLM-generated commonsense and VTKG-C (Lee et al., 2023), an existing multimodal physical knowledge graph.

For LLM-generated commonsnese, we prompt InternVL8B to generate commonsense knowledge phrases about objects, people, relationships, behaviors, and the contextual background of an input image, using 100 samples per task, and inject this in the same setup as MMCOMET. Figure 9 (Appendix H.4) highlights the effectiveness of MMCOMET's multimodal commonsense knowledge for reasoning-dominated tasks. In particular, MMCOMET outperforms all other setups for VQA and VCR, where model reasoning benefits from the integrated physical, social, and eventive knowledge. Notably, LLM-generated knowledge performs worst across all tasks, indicating that the structured knowledge from the commonsense graphs better directs visual grounding than naive knowledge integration. Perception-dominated tasks such as IC and IE benefit less from commonsense knowledge integration; however, greater performance gains are still achieved over text-only knowledge, demonstrating more robust knowledge. More detailed results and comparison to VTKG-C are provided in the Appendix H.4

6 Knowledge Utility Estimation and Selection with MMCOMET

Naively integrating all available external knowledge can be harmful when the retrieved knowledge is irrelevant to an instance. We therefore propose an adaptive knowledge router that selects either one of the available MMCOMET knowledge routes or a no-knowledge route. The router predicts an instance-level utility score based on the estimated change in prediction margin relative to the baseline. It is trained to preserve routes associated with positive gains while bypassing knowledge when predicted utility is insufficient. Thus, the router treats knowledge integration as a selective decision rather than a fixed augmentation.

Table 8: Adaptive Router (R) on MMStar against baseline and full MMCOMET (M). MMCOMET with a router generally improves upon the baseline or mitigates degradation.

Model	Setup	Overall	LR	CP	MA
Qwen32B	Baseline	0.7160	0.6912	0.7614	0.7500
	M (Full)	0.7000	0.7206	0.7500	0.7024
	M + R	**0.7260**	**0.7353**	**0.7841**	**0.7619**
LLaVa34B	Baseline	0.5060	0.4265	**0.6364**	**0.4881**
	M (Full)	0.4720	0.4706	0.6023	0.4405
	M + R	**0.5080**	**0.5000**	**0.6364**	0.4762
Mistral24B	Baseline	**0.6020**	0.5735	0.6818	0.6071
	M (Full)	0.5840	0.5441	0.6932	0.5952
	M + R	**0.6020**	**0.5882**	**0.7045**	**0.6310**
Qwen8B	Baseline	0.6460	0.6176	**0.7386**	0.6548
	M (Full)	0.6300	0.6029	0.7045	0.6548
	M + R	**0.6520**	**0.6324**	0.7159	**0.6905**

We evaluate the router on MMStar (Chen et al., 2024), a comprehensive VQA benchmark with six heterogeneous question categories. The candidate set contains seven MMCOMET routes, comprising individual commonsense domains and their combinations. This setting tests whether selective integration retains useful knowledge while avoiding the degradation caused by indiscriminate augmentation. Table 8 shows that applying all MMCOMET knowledge uniformly (*M (Full)*) substantially reduces performance relative to the baseline. In contrast, MMCOMET with the adaptive router (*M + R*) consistently improves overall performance over both the baseline and full knowledge integration. The largest gains occur on Logical Reasoning (*LR*), particularly for larger models, indicating that selectively retrieved visually grounded commonsense knowledge is beneficial for reasoning-intensive questions. Math (*MA*) and Coarse Perception (*CP*) also improved from *M (Full)*, showing that the router mitigates degradation from fixed knowledge routes. These results indicate that external commonsense knowledge utility is heterogeneous across inputs and question categories. Selective routing is more robust, retaining beneficial routes when estimated utility is positive and defaulting to the no-knowledge route when augmentation is unhelpful. When applied to the VCR and VQAv2 datasets using 5-fold CV, routed MMCOMET outperforms the baseline by up to +0.0010 and +0.0114, respectively. More details for the router may be found in Appendix K.

7 Conclusion

We introduced MMCOMET, a large-scale multimodal commonsense knowledge graph that grounds physical, social, and eventive knowledge with representative visual evidence. By extending ATOMIC2020 through a scalable hybrid alignment pipeline, MMCOMET provides nearly one million multimodal commonsense tuples for retrieval-augmented vision-language reasoning and generation. Experiments across diverse downstream tasks show that visually grounded commonsense consistently improves over text-only and LLM-generated knowledge, particularly in reasoning-intensive settings. We further propose an adaptive knowledge router for heterogeneous knowledge requirements, enabling dynamic route selection with a no-knowledge fallback when no benefit from commonsense knowledge is expected. Overall, MMCOMET establishes a reusable resource and benchmark for studying grounded commonsense integration in multimodal systems.

AI USE STATEMENT

In this work, generative AI tools were used only to edit the manuscript for grammar and readability. We have not used generative AI tools for any of the required-disclosure tasks. All such work is the authors' own. All AI-assisted edits were reviewed by authors. We take full responsibility for the final content of this work, including text, claims or artifacts.

ETHICS STATEMENT

Further details including licensing, risk and institutional review are provided in Appendix B

Limitations First, our retrieval pipeline relies on CLIP ViT-B/32 embeddings, which, while comparatively dated, reflect a deliberate scalability/efficiency trade-off at the scale of our benchmark ($\sim$3.9M images); adopting a stronger encoder is left for future work. Second, MMCOMET is constructed from web-crawled images, which raises ethical considerations around data provenance, consent, and the potential presence of sensitive or inappropriate content that automated crawling cannot fully filter out; we discuss our mitigation efforts (e.g., filtering pipelines) in the ethics statement, and note that source URL expiration over time may also affect long-term reproducibility, which we mitigate by releasing the full precomputed image feature matrix alongside the benchmark. Third, we fix the number of retrieved phrases k to a single, empirically stable value across all tasks rather than tuning it per task to keep the framework simple and broadly applicable; task-specific retrieval depths (e.g., for VQA, VCR, or captioning) may yield further gains, which we leave to future work.

REFERENCES

Houda Alberts, Ningyuan Huang, Yash Deshpande, Yibo Liu, Kyunghyun Cho, Clara Vania, and Iacer Calixto. VisualSem: a high-quality knowledge graph for vision and language. In Duygu Ataman, Alexandra Birch, Alexis Conneau, Orhan Firat, Sebastian Ruder, and Gozde Gul Sahin (eds.), *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pp. 138–152, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.mrl-1.13. URL https://aclanthology.org/2021.mrl-1.13/.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (eds.), *Computer Vision – ECCV 2016*, pp. 382–398, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46454-1.

Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 2014.

Hua Cai, Xuli Shen, Qing Xu, Weilin Shen, Xiaomei Wang, Weifeng Ge, Xiaoqing Zheng, and Xiangyang Xue. Improving empathetic dialogue generation by dynamically infusing commonsense knowledge. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 7858–7873, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.498. URL https://aclanthology.org/2023.findings-acl.498/.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

Sebastián Ferrada, Benjamin Bustos, and Aidan Hogan. Imgpedia: A linked dataset with content-based analysis of wikimedia images. In Claudia d'Amato, Miriam Fernandez, Valentina Tamma, Freddy Lecue, Philippe Cudré-Mauroux, Juan Sequeda, Christoph Lange, and Jeff Heflin (eds.),

The Semantic Web – ISWC 2017, pp. 84–93, Cham, 2017. Springer International Publishing. ISBN 978-3-319-68204-4.

François Gardères, Maryam Ziaeefard, Baptiste Abeloos, and Freddy Lecue. ConceptBert: Concept-aware representation for visual question answering. In Trevor Cohn, Yulan He, and Yang Liu (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 489–498, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.44. URL https://aclanthology.org/2020.findings-emnlp.44/.

Biao Gong, Shuai Tan, Yutong Feng, Xiaoying Xie, Yuyuan Li, Chaochao Chen, Kecheng Zheng, Yujun Shen, and Deli Zhao. Uknow: A unified knowledge protocol with multimodal knowledge graph datasets for reasoning and vision-language pre-training. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 9612–9633. Curran Associates, Inc., 2024. doi: 10.52202/079017-0306. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/126784e4d5a92afff92d13aee155554b-Paper-Datasets_and_Benchmarks_Track.pdf.

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 *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6325–6334, 2017. doi: 10.1109/CVPR.2017.670.

Jian Guan and Minlie Huang. UNION: An Unreferenced Metric for Evaluating Open-ended Story Generation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9157–9166, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.736. URL https://aclanthology.org/2020.emnlp-main.736/.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. Realm: retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

MD Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CsUR)*, 2019.

Ting-Hao Kenneth Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. Visual storytelling. In Kevin Knight, Ani Nenkova, and Owen Rambow (eds.), *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1233–1239, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1147. URL https://aclanthology.org/N16-1147/.

Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. Comet-atomic 2020: On symbolic and neural commonsense knowledge graphs. In *AAAI*, 2021.

Xuan Ju, Ailing Zeng, Yuxuan Bian, Shaoteng Liu, and Qiang Xu. Pnp inversion: Boosting diffusion-based editing with 3 lines of code. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 23395–23422, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/661caac7729aa7d8c6b8ac0d39ccbc6a-Paper-Conference.pdf.

Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):664–676, 2017. doi: 10.1109/TPAMI.2016.2598339.

Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5583–5594. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/kim21k.html.

Jaejun Lee, Chanyoung Chung, Hochang Lee, Sungho Jo, and Joyce Jiyoung Whang. VISTA: Visual-textual knowledge graph representation learning. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 7314–7328, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.488. URL https://aclanthology.org/2023.findings-emnlp.488/.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision – ECCV 2014*, pp. 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916. Curran Associates, Inc., 2023. doi: 10.52202/075280-1516. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf.

Hugo Liu and Push Singh. Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, 2004.

Ye Liu, Hui Li, Alberto Garcia-Duran, Mathias Niepert, Daniel Onoro-Rubio, and David S. Rosenblum. Mmkg: Multi-modal knowledge graphs. In Pascal Hitzler, Miriam Fernández, Krzysztof Janowicz, Amrapali Zaveri, Alasdair J.G. Gray, Vanessa Lopez, Armin Haller, and Karl Hammar (eds.), *The Semantic Web*, pp. 459–474, Cham, 2019. Springer International Publishing. ISBN 978-3-030-21348-0.

Tong Lu, Zhichun Wang, Yaoyu Zhou, Yiming Guan, Zhiyong Bai, and Junsheng Du. Scimkg: A multimodal knowledge graph for science education with text, image, video and audio. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(18):15466–15474, Mar. 2026. doi: 10.1609/aaai.v40i18.38574. URL https://ojs.aaai.org/index.php/AAAI/article/view/38574.

Yubo Ma, Zehao Wang, Mukai Li, Yixin Cao, Meiqi Chen, Xinze Li, Wenqi Sun, Kunquan Deng, Kun Wang, Aixin Sun, and Jing Shao. MMEKG: Multi-modal event knowledge graph towards universal representation across modalities. In Valerio Basile, Zornitsa Kozareva, and Sanja Stajner (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 231–239, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-demo.23. URL https://aclanthology.org/2022.acl-demo.23/.

Oscar Mañas, Benno Krojer, and Aishwarya Agrawal. Improving automatic vqa evaluation using large language models. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press, 2024. ISBN 978-1-57735-887-9. doi: 10.1609/aaai.v38i5.28212. URL https://doi.org/10.1609/aaai.v38i5.28212.

Tuan-Phong Nguyen, Simon Razniewski, Julien Romero, and Gerhard Weikum. Refined commonsense knowledge from large-scale web contents. *IEEE Transactions on Knowledge and Data Engineering*, 35(8):8431–8447, 2023. doi: 10.1109/TKDE.2022.3206505.

Daniel Oñoro-Rubio, Mathias Niepert, Alberto García-Durán, Roberto González-Sánchez, and Roberto J. López-Sastre. Answering visual-relational queries in web-extracted knowledge graphs. In *Automated Knowledge Base Construction (AKBC)*, 2019. URL https://openreview.net/forum?id=BylEpe9ppX.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Simon Razniewski, Hiba Arnaout, Shrestha Ghosh, and Fabian Suchanek. Completeness, recall, and negation in open-world knowledge bases: A survey. *ACM Comput. Surv.*, 56(6), February 2024. ISSN 0360-0300. doi: 10.1145/3639563. URL https://doi.org/10.1145/3639563.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. Atomic: an atlas of machine commonsense for if-then reasoning. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press, 2019. ISBN 978-1-57735-809-1. doi: 10.1609/aaai.v33i01.33013027. URL https://doi.org/10.1609/aaai.v33i01.33013027.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. BLEURT: Learning robust metrics for text generation. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7881–7892, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.704. URL https://aclanthology.org/2020.acl-main.704/.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2556–2565, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1238. URL https://aclanthology.org/P18-1238/.

Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: an open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, pp. 4444–4451. AAAI Press, 2017.

Niket Tandon, Gerard de Melo, Fabian Suchanek, and Gerhard Weikum. Webchild: harvesting and organizing commonsense knowledge from the web. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, WSDM '14, pp. 523–532, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450323512. doi: 10.1145/2556195.2556245. URL https://doi.org/10.1145/2556195.2556245.

Eileen Wang, Caren Han, and Josiah Poon. Rovist: Learning robust metrics for visual storytelling. In *Findings of the Association for Computational Linguistics*, 2022.

Eileen Wang, Soyeon Caren Han, and Josiah Poon. Sco-vist: Social interaction commonsense knowledge-based visual storytelling. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, 2024.

Meng Wang, Haofen Wang, Guilin Qi, and Qiushuo Zheng. Richpedia: A large-scale, comprehensive multi-modal knowledge graph. *Big Data Research*, 22:100159, 2020. ISSN 2214-5796. doi: https://doi.org/10.1016/j.bdr.2020.100159. URL https://www.sciencedirect.com/science/article/pii/S2214579620300277.

Xin Wang, Benyuan Meng, Hong Chen, Yuan Meng, Ke Lv, and Wenwu Zhu. Tiva-kg: A multimodal knowledge graph with text, image, video and audio. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, pp. 2391–2399, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701085. doi: 10.1145/3581783.3612266. URL https://doi.org/10.1145/3581783.3612266.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Yinan Wu, Xiaowei Wu, Junwen Li, Yue Zhang, Haofen Wang, Wen Du, Zhidong He, Jingping Liu, and Tong Ruan. Mmpedia: A large-scale multi-modal knowledge graph. In Terry R. Payne, Valentina Presutti, Guilin Qi, María Poveda-Villalón, Giorgos Stoilos, Laura Hollink, Zoi Kaoudi, Gong Cheng, and Juanzi Li (eds.), *The Semantic Web – ISWC 2023*, pp. 18–37, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-47243-5.

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

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6713–6724, 2019. doi: 10.1109/CVPR.2019.00688.

Hongming Zhang, Daniel Khashabi, Yangqiu Song, and Dan Roth. Transomcs: From linguistic graphs to commonsense knowledge. In Christian Bessiere (ed.), *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pp. 4004–4010. International Joint Conferences on Artificial Intelligence Organization, 7 2020. doi: 10.24963/ijcai.2020/554. URL https://doi.org/10.24963/ijcai.2020/554. Main track.

Jingdan Zhang, Jiaan Wang, Xiaodan Wang, Zhixu Li, and Yanghua Xiao. Aspectmmkg: A multimodal knowledge graph with aspect-aware entities. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, pp. 3361–3370, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701245. doi: 10.1145/3583780.3614782. URL https://doi.org/10.1145/3583780.3614782.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 563–578, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1053. URL https://aclanthology.org/D19-1053/.

Yimin Zhou, Yiwei Sun, and Vasant Honavar. Improving image captioning by leveraging knowledge graphs. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 283–293, 2019. doi: 10.1109/WACV.2019.00036.

Category	Subcategory	Location	Key Items (Purpose)
Construction	***How is MMCOMET constructed?***		
	Base Commonsense Topology	Sec. 3.1	Table 1 (MMKG Comparison), Fig. 2 (Construction Pipeline Overview)
	Construction Pipeline	Sec. 3.2, Sec. 3.3	-
	Graph Statistics	App. D.1, App. D.2, App. D.3	Table 10 (Relation Definitions), Table 11 (Web-based retrieval Relations), Table 12 (Graph Statistics)
Internal Analysis	***Is MMCOMET visually well-grounded?***		
	Annotation Filtering	Sec. 4, App. E.1	Table 2 (Full/Partial Match Rates by Relation)
	Concreteness & Alignment Analysis	Sec. 4, App. J.2, App. J.3	Fig. 15, Fig. 16
Downstream Task	***Does visual grounding help downstream tasks?***		
	Evaluation Setup & Reproducibility	App. C, App. F	Table 13, 14 (Evaluation Setup), Fig 4 (VST Setup), Table 21 - 31 (Prompts)
	Downstream Task Result	Sec. 5	Table 3 – 7 (Main Result; VCR, VQA, VST, IE, IC), Table 15
	Qualitative Analysis	App. I	Fig. 10 – Fig. 14 (Qualitative Analysis)
Knowledge Router	***Which knowledge actually drives the gains – and how should it be selected?***		
	Effect of Knowledge Source	Sec. 5.7	Fig. 9 (Result by Knowledge Source), Table 18 (Comparision with VTKG-C)
	Domain Coverage (Physi-cal/Social/Eventive)	App. H.2	Table 16(VCR, VQA, IC)
	Alignment Strategy Ablation	App. H.3	Table 17 (VCR & Image Captioning by image source)
	Breaking Point	App. H.1	Fig. 3 (Category Breakdown), Fig. 6 (Ambiguity vs. Accuracy), Fig. 7, 8 (Performance Gap by Category)
	Router Experiment	Sec. 6, App. K	Table 8 (Main Result), Table 20 (Full Result), Alg. 1, 2 (Algorithm)
Ethical Considerations	*What ethical safeguards are in place?*		
	Released Artifacts & Licensing	App. B	-
	Privacy & Social-Inferential Risk	App. B	-
	Sensitive Content & Safety Filtering	App. B, App. D.4	Fig. 5 (SafeSearch Filtering, Manual Spot-check on 10K Images)
	Bias & Stereotyping	App. B	-

Table 9: Structural overview of MMCOMET.

A Overview

This Appendix provides additional details for dataset construction, evaluation setup, extended results, and qualitative examples supporting the main paper. In particular, we organize our extended experimental analyses around three research questions:

- **RQ1: Does multimodal grounding matter?** Text-only knowledge bases cannot resolve which of several plausible commonsense associations applies to a specific visual scene, so we test whether grounding commonsense in images resolves this ambiguity. We compare MMCOMET against text-only knowledge (TextKG), LLM-generated knowledge, and an existing multimodal commonsense knowledge graph (VTKG-C) in Sec. 5.7 and App. H.4. MMCOMET outperforms all three alternatives in most settings, and the margin is largest on reasoning-heavy tasks. This suggests that visual grounding supplies information that text-only, LLM-generated, or single-domian knowledge does not reliably recover.
- **RQ2: Which components of MMCOMET matter?** MMCOMET's contribution combines two design choices: relation coverage beyond the physical domain, and a hybrid alignment strategy for linking phrases to images. We ablate each in Appendix H.2 and Appendix H.3. The domains that matter are task-dependent. Removing social and eventive coverage disproportionately hurts VCR, while replacing the hybrid alignment with a single-signal alternative degrades retrieval precision, confirming both components are necessary. Broad coverage therefore matters more than any single task-specific optimum
- **RQ3: Which knowledge actually drives the gains – and how should it be selected?** If MMCOMET's benefit reflects commonsense reasoning rather than a uniform boost, it should concentrate where reasoning is required. We analyze gains by sample ambiguity on VCR (App. H.1) and by question category on MMStar (Sec. 6, App. K). The gap between MMCOMET and TextKG widens with ambiguity, and on MMStar full injection helps Logical Reasoning for larger models but hurts perception-heavy categories and overall accuracy. This motivates an adaptive utility router that selects a knowledge route or falls back to baseline, which recovers or slightly exceeds the baseline. We therefore view it as evidence for selective knowledge use rather than a large accuracy gain.

B Ethics Statement

This work proposes a resource that grounds text-only commonsense knowledge with visual evidence, most of which is retrieved via web search. We acknowledge that this raises additional concerns, which we attempt to address in this section.

Released Artifacts and Licensing. For all images, we release only the final visual representations used to construct the resource, along with the complete triplet-image association and embedding-extraction metadata. We do not redistribute raw images, whether web-sourced or from other open-source datasets. For web-retrieved images, we further release provenance metadata and licensing details. We do not include URLs to respect copyright and privacy constraints and to prevent potential reverse identification.

Privacy and Social-Inferential Risk. We acknowledge that some images used may contain identifiable individuals that has been systematically paired with inferred attributes, intents, and emotional states. We attempt to mitigate this by not releasing the full raw images however, we don't have full control for images from open-source datasets. Instead, we provide a clear take down mechanism for any image or triplet-image association upon request to the corresponding author or through the released repository.

Sensitive Content and Safety Filtering. For all web-retrieved images, we used the Google Custom Search API with active SafeSearch enabled. This provides us with robust safety filtering against sensitive and explicit content as illustrated in Figure 5 and discussed in Appendix D.4. However, we acknowledge that enabling this option is not 100% perfect. As such, we conducted random spot-checking and manual evaluation on 10k images and 1000 phrases.

Bias and Stereotyping. MMCOMET uses naturally occurring web imagery to preserve the diversity of real-world visual contexts rather than relying only on curated benchmark images. At the same time, we acknowledge that real-world web data can reproduce existing social and demographic biases, including occupational, gender, and racial stereotypes.

Institutional Review Board Statement. The study was approved by the Institutional Review Board (Exemption Number: PIRB-2025-E020, Date: June 9, 2025).

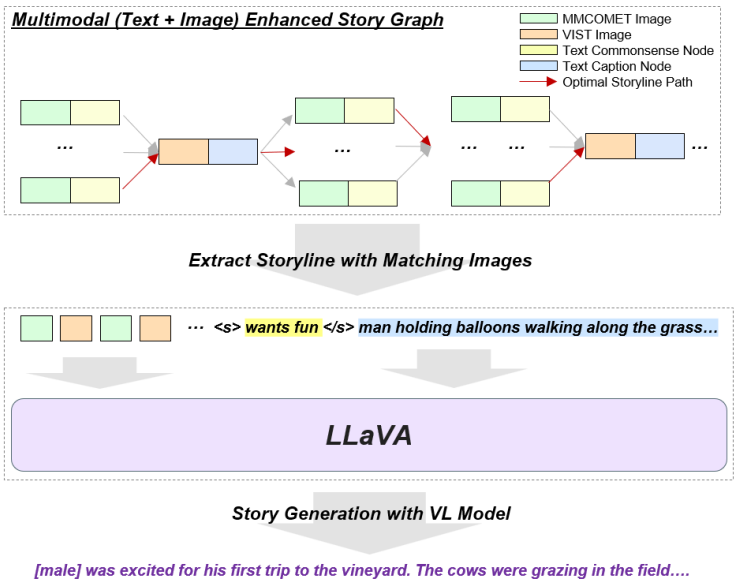


Figure 4: SCO-VIST framework enhanced with MMCOMET.

C Multimodal Commonsense Enhanced Visual Storytelling

Figure 4 depicts how SCO-VIST's storygraph is enhanced with information from MMCOMET. In SCO-VIST, each node in the graph only has a textual modality (e.g. a caption description or commonsense phrase). We incorporate image features into the story graph by using the generated phrase to query MMCOMET to find the matching image. For simplicity, if the phrase is not found in MMCOMET, we delete the node from the story graph. The CLIP features of the top-found image are then concatenated with the textual features for the commonsense nodes in the story graph, while for the caption nodes, we directly use the image features from the image sequence prompt. The story graph edge weights are then recalculated using the cosine similarity between adjacent nodes. The storyline is then extracted from the graph using a shortest-path search algorithm. The textual storyline, along with the MMCOMET and VIST images, is finally fed into a vision-language model (LLaVA) for story generation. Additionally, two integration setups are tested. LLaVA+MMCOMET^{S} enriches the story graph with MMCOMET image embeddings, and the input into the LLaVA story decoder is the storyline obtained from the multimodal story graph + 5 VIST images. LLaVA+MMCOMET^{S+I} additionally feeds the MMCOMET images, the enriched storyline, and the 5 VIST images.

D Multimodal Commonsense Graph Analysis

D.1 Comet-ATOMIC2020 Relation Types

Table 10 shows the different relations and their definitions, as well as corresponding examples in the original Comet-ATOMIC2020 graph. Green, orange and blue highlighting indicates social, event and physical relations respectively. Note that for simplicity, in this paper, we refer to `xEffect` and `oEffect` as `Effect`, `xReact` and `oReact` as `React`, and `xWant` and `oWant` as `Want`. The `x` prepended in front of the relation name is also removed for consistency e.g. `xNeed` is referred to as `Need`.

D.2 Web Retrieval Relation Type Analysis

Table 11 reports the proportion of head and tail phrases routed to web-based retrieval under the concreteness threshold. Consistent with their lower concreteness scores, social and eventive relations require substantially more web searches than physical relations. An interesting asymmetry emerges within certain physical relations such as `HasProperty` and `Desires/NotDesires`. While head phrases typically contain tangible noun entities and therefore exhibit high concreteness, tail phrases often describe abstract properties or mental states. As a result, tail phrases in these relations show a disproportionately high reliance on web-based grounding. The analysis confirms a

Table 10: Different Relations in Comet-ATOMIC2020 and corresponding examples. Examples and definitions taken from the official Comet-ATOMIC2020 paper (Hwang et al., 2021).

Relation	Description	Example (<head, relation, tail>)
ObjectUse	describes everyday affordances or uses of objects	bread ObjectUse make french toast
AtLocation	spatially describes the where an entity is likely to be found	bread AtLocation pantry
MadeUpOf	describes a part, portion or makeup of an entity	bread MadeUpOf dough
HasProperty	describes entities' general characteristics	bread HasProperty nice to eat
CapableOf	describe abilities and capabilities of everyday living entities	baker CapableOf coat cake with icing
Desires	desire of sentient entity	baker Desires quality ingredients
NotDesires	non-desire of sentient entity	baker NotDesires bad yeast
isAfter	events that can follow an event	X runs out of steam isAfter X exercises in the gym
HasSubEvent	provides the internal structure of an event	X runs out of steam HasSubEvent become tired
isBefore	events that can precede an event	X runs out of steam isBefore X hits the showers
HinderedBy	hindrances that obstruct the natural path of an event	X runs out of steam HinderedBy drinks too much coffee
Causes	causal relation between two events or entities	X runs out of steam Causes takes a break
xReason	provides a post-fact explanation of the cause of an event	X runs out of steam xReason did not eat breakfast
xNeed	describes a precondition for an agent to achieve the event	X runs out of steam xNeed do something tiring
xAttr	describes personas or attributes perceived by others given an event	X runs out of steam xAttr old
xEffect	actions that happen to agent X that may occur after the event	X runs out of steam xEffect drinks some water
oEffect	actions that happen to agent Y that may occur after the event	X votes for Y oEffect receives praise
xReact	emotional reactions of agent X in an event	X runs out of steam xReact tired
oReact	emotional reactions of agent Y in an event	X votes for Y oReact grateful
xWant	agent X's postcondition desires after an event	X runs out of steam xWant to get some energy
oWant	agent Y's postcondition desires after an event	X votes for Y oWant thank X
xIntent	defines the likely intent of an agent	X votes for Y xIntent to give support

strong correlation between relation-level concreteness and retrieval routing behavior. These findings empirically justify the concreteness-aware hybrid alignment strategy, as a uniform similarity-based approach would inadequately ground abstract social and eventive commonsense expressions.

Table 11: The left sub-table shows the average concreteness scores for each commonsense relation (sample size in brackets) in ascending order. The second and third sub-tables show the proportion of heads and tail phrases across each relation that required web search. Green, orange and blue highlighting indicate social, event and physical relations, respectively.

Relation	Avg. Concreteness
React (94,228)	2.95
Attr (115,815)	2.98
Causes (376)	3.29
Intent (49,312)	3.40
Effect (113,494)	3.53
Reason (334)	3.55
HasSubEvent (12,845)	3.58
Want (152,989)	3.58
Need (89,391)	3.65
HinderedBy (106,647)	3.80
HasProperty (5,617)	3.80
Desires (2,737)	3.88
ObjectUse (165,590)	3.90
isBefore (23,208)	3.91
NotDesires (2,838)	3.91
isAfter (22,453)	3.94
CapableOf (7,968)	3.96
MadeUpOf (3,345)	4.08
AtLocation (20,221)	4.15

Relation	% Heads Searched
NotDesires	2.82
Desires	2.85
AtLocation	18.78
MadeUpOf	20.03
HasProperty	22.77
CapableOf	23.98
ObjectUse	25.26
Reason	50.90
Intent	51.76
Need	54.11
React	54.85
Effect	54.85
isAfter	55.96
isBefore	56.07
Want	56.15
HinderedBy	56.19
Attr	57.13
HasSubEvent	60.19
Causes	65.96

Relation	% Tails Searched
AtLocation	25.52
MadeUpOf	36.47
isAfter	48.98
isBefore	53.50
CapableOf	59.55
HinderedBy	64.72
ObjectUse	67.65
HasProperty	71.46
HasSubEvent	73.03
NotDesires	74.03
Need	75.94
Reason	76.35
Effect	76.54
Causes	77.93
Desires	79.65
Want	80.92
Intent	90.35
Attr	98.82
React	99.48

D.3 Graph Statistics

Table 12 shows the number of unique nodes, total edges, unique web searched images, unique images acquired from open-source datasets and total unique images for each relation, and across all relations.

D.4 Error Analysis

We analyse common errors that occurred in the process of finding matching images for the commonsense phrase. The 3 main types of errors and visual examples are depicted in Figure 5. One challenging issue that remains is finding matching images for non-visually grounding commonsense. For instance, the first 2 commonsense examples in Figure 5, '*rhetoric*' and '*leap year*' are very abstract terms and hence, difficult to visually ground even when employing web search. Consequently, the images obtained usually contains text or quotes as with the case for the matched images found

for '*rhetoric*'. In other cases, the retrieved images may sometimes be related to movies or media where the title of the media is similar to the commonsense phrase, such as with the case of '*leap year*,' which seems to have retrieved images related to a movie that has the same name as the commonsense phrase. The second issue found was that sometimes the relationship between people and concepts in the commonsense phrase is not sufficiently captured by the image. This typically occurs when the phrase mentions two people - '*PersonX*'and '*PersonY*'. For instance, the matching images found from '*PersonX loves PersonY's car*' either solely focus on the idea of 'love' as in the left image or the concept of 'car' as in the image on the right. However, the complex relationship between the first concept of a 'person loving' and the second concept, 'another person's car', is much more difficult to capture. Finally, there are cases where the commonsense phrase mentions violent or inappropriate topics. Inevitably, these images are most likely blocked by Google's web search engine, such that the images obtained may not fully match the search phrase. For example, for the commonsense phrase '*commit murder*', we again obtained images related to movies/television shows as shown in the left image. At best, we tend to obtain images that are partial matches, such as in the image on the right, where the action is not explicitly shown but is implied.

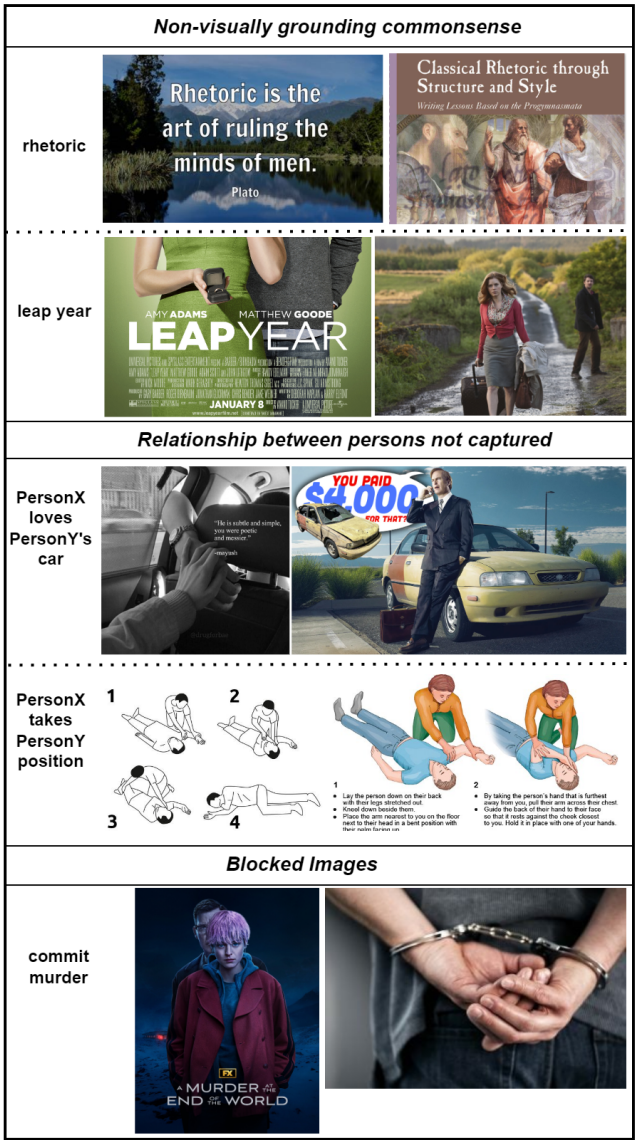


Figure 5: Examples of retrieved images for 3 different error cases found during the image collection stage.

E HUMAN EVALUATION INFORMATION

E.1 INTRINSIC EVALUATION

We conducted our annotation process with three annotators: a PhD student, a Postdoctoral researcher, and an Associate Professor. These annotators were recruited based on their expertise in the relevant research to ensure high-quality annotations. Our annotators come from diverse geographical backgrounds—Asia, Australia, and Europe—reflecting a broad perspective in the annotation process. The human evaluation was conducted on a voluntary basis, and no monetary compensation was provided. Before participation, we provided clear instructions explaining the purpose of the evaluation and how the collected data would be used. The annotators were informed that their feedback would contribute to the assessment and refinement of our system's performance, and they participated with full awareness and consent.

Table 12: Graph statistics for MMCOMET showing total unique nodes, edges, web-searched images, open-source dataset images and total images across the different relations.

Relation	# Nodes	# Edges	# WSE Images	# OID Images	# Total Images
ObjectUse	137,144	165,590	767,124	242,211	1,009,335
AtLocation	11,216	20,221	35,204	62,730	97,934
MadeUpOf	3,464	3,345	11,047	24,651	35,698
HasProperty	6,123	5,617	32,233	35,738	67,971
CapableOf	8,872	7,968	45,005	44,675	89,680
Desires	2,832	2,737	18,750	14,911	33,661
NotDesires	2,991	2,838	18,312	15,602	33,914
isAfter	36,181	22,453	162,312	97,386	259,698
HasSubEvent	10,670	12,845	67,522	42,541	110,063
isBefore	37,482	23,208	175,957	96,981	272,938
HinderedBy	103,844	106,647	588,983	164,975	753,958
Causes	531	376	3,299	3,471	6,770
Reason	441	334	2,622	2,888	5,510
Need	80,575	89,391	490,672	136,668	627,340
Attr	29,650	115,815	175,477	73,072	248,549
Effect	90,983	113,494	567,801	134,982	702,783
React	35,537	94,228	228,856	77,635	306,491
Want	107,268	152,989	689,798	146,314	836,112
Intent	45,771	49,312	292,494	92,549	385,043
All	571,404	989,408	3,476,098	435,695	3,911,793

E.2 EXTRINSIC EVALUATION

We conducted our annotation process with eleven annotators, including one undergraduate student, two graduated assistants, four Master's students, three PhD students, and one Postdoctoral researcher. These annotators were recruited based on their familiarity with the task and relevant research experience to ensure high-quality annotations. All annotators are from Asia, providing a consistent cultural background for the evaluation process. The group consisted of five male and six female annotators, ensuring gender balance. Human evaluation was conducted on a voluntary basis and no monetary compensation was provided. Before participation, we provided clear instructions explaining the purpose of the evaluation and how the collected data would be used. The annotators were informed that their feedback would contribute to the assessment and refinement of our system's performance, and they participated with full awareness and consent.

Each annotator evaluated 80 samples drawn from four tasks that we evaluated (VCR, VQA, Image Editing, Image Captioning), with 20 samples per task. For each sample, annotators were presented with the input image, the model outputs under each setting, and the retrieved commonsense knowledge phrases provided to model. They rated two statements on a 4-point Likert scale (1 = Strongly Disagree, 4 = Strongly Agree), (1) *"Is the retrieved commonsense knowledge relevant to the image and input?"* and (2) *"Does the knowledge help improve reasoning or inference for this task?"*.

The average scores across all tasks were 3.1 and 2.9 for Q1 and Q2, respectively, both exceeding the neutral midpoint of 2.5. These results suggest that annotators generally found the retrieved

Table 14: Task dataset statistics, metrics, and models used.

Task	Dataset	#Items	Metric	Models
Visual Commonsense Reasoning	VCR (val)	26,534	**Accuracy**	Qwen3-VL-8B-Instruct, Qwen3-VL-32B-Instruct, InternVL3-38B-hf
Visual QA	VQAv2 (val)	214,354	**Accuracy** *(VQA-style soft scoring; a predicted answer receives partial credit based on agreement with human annotations)*	Qwen3-VL-8B-Instruct, gemma-4-31B-it
Image Editing	PIE-Bench	700	**PSNR** *(pixel-level distortion)*, **SSIM** *(structural similarity)*, **LPIPS** *(perceptual similarity)*	Qwen-Image-Edit
Image Captioning	Flickr30K (1K Test set)	1,000	**BLEU** *(n-gram precision)*, **CIDEr** *(consensus-based similarity)*, **ROUGE-L** *(longest common subsequence)*	Qwen3-VL-8B-Instruct, Qwen3-VL-32B-Instruct, gemma-4-31B-it
Visual Story Telling	VIST	Train: 8,031 Val: 998 Test: 1,011	**RoViST** *(visual grounding, language coherence, repetition)*, **Perplexity** *(NLG metric)*, **UNION** *(NLG metric)*, **SPICE** *(semantic similarity)*, **BLEURT** *(semantic similarity)*, **MoverScore** *(semantic similarity)*	SCOVIST+LLAVA

commonsense knowledge to be relevant and beneficial, even when evaluated across both success and fail cases. Notably, even in cases where MMCOMET did not outperform the Baseline on automatic metrics, annotators still perceived the retrieved knowledge as helpful for reasoning, highlighting a potential gap between automatic evaluation metrics and human judgement.

F Evaluation Setup

Table 14 shows the dataset statistics, metric descriptions, and primary models tested for each task. We provide the relevant hyperparameters and details in Table 13. For all tasks except Visual Story Telling, we compare three conditions: **Baseline** (no knowledge-injected), **TextKG** (text-based knowledge injection), and **MMCOMET** (multimodal knowledge injection). All experiments were conducted with single H100 GPU with 94GB GPU memory.

Table 13: Relevant hyperparameters.

Hyperparameter	Value
Common	
CLIP backbone	ViT-B/32
Candidate images (n)	15
Retrieved phrases (k)	5
Random seed	42
Max Tokens	
Max Tokens - VCR	10
Max Tokens - VQA	10
Max Tokens - Captioning	64
Image Editing	
Inference steps	50
CFG scale	4.0
Retrieved phrases (k)	1

For the knowledge retrieval we tried several methods including (1) **L2 retrieval**, which retrieves the top-n visually similar images and re-ranks associated phrases by $s = \|E_{\text{img}} - E_{\text{phrase}}\|_2 + \|E_{\text{query}} - E_{\text{phrase}}\|_2$, (2) **FAISS-based direct search**, which retrieves phrases directly from the full phrase index using a combined embedding $E_{\text{combined}} = (E_{\text{img}} + E_{\text{query}})/2$, (3) **Weighted scoring**, which computes $s = w_{i2i} \cdot \text{sim}(E_{\text{img}}, E^{\text{KG}}_{\text{img}}) + w_{i2t} \cdot \text{sim}(E_{\text{img}}, E_{\text{phrase}}) + w_{t2t} \cdot \text{sim}(E_{\text{query}}, E_{\text{phrase}})$ and (4) **MMR**, which promotes diversity by $\text{MMR} = \lambda \cdot s_{\text{rel}} - (1-\lambda) \cdot \max_{p_j \in S} \text{sim}(p_i, p_j)$, where S is the set of selected phrases. As performance differences across methods were negligible, making the choice of retrieval method a minor implementation detail rather than a critical design choice. Consequently, we adopt L2 retrieval for all experiments due to its simplicity.

Regarding knowledge injection, **TextKG** inserts k KG triplets, which are combinations of 2 phrases, retrieved via text query into the prompt. **MMCOMET** inserts k commonsense phrases from both head and tail entities, yielding $2k$ phrases per sample. We empirically set $k = 5$, as it provided consistently strong performance across tasks (Image Editing shows consistently strong when $k = 1$.). Exploring task-specific optimal values of k remains a direction for future work.

VCR VCR[3] (Zellers et al., 2019) is a visual commonsense reasoning benchmark that originally requires models to select a correct answer given a question (Q→ A) or to select a rationale given a question-answer pair (QA→ R). In this work, we intentionally modify the task setup to require models to select a rationale given only the question (Q→ R), without access to the ground-truth answer. This modification is designed to prevent shortcut learning, where a model could otherwise exploit the provided answer to trivially infer the rationale. The prompt template used for this task is shown in Table 21 - 23.

VQA VQAv2[4] (Goyal et al., 2017) is a multimodal reasoning benchmark that asks open-ended questions about images. The questions require an understanding of vision, language and commonsense knowledge to answer. The prompt template used for this task is shown in Table 24 – 26.

Image Captioning Flickr30K[5] (Young et al., 2014) is a widely used image captioning benchmark. We evaluate on the Karpathy test split (Karpathy & Fei-Fei, 2017), which is the standard evaluation protocol for image captioning tasks. The prompt template used for this task is shown in Table 29 – 31.

Image Editing PIE_Bench[6] (Ju et al., 2024) is a benchmark for evaluating text-guided image editing models, consisting of 700 images across 10 editing categories. We evaluate edited image quality using PSNR, SSIM, and LPIPS, which measure pixel-level fidelity, perceptual structural similarity, and learned perceptual distance, respectively. The prompt template used for this task is shown in Table 27 and 28.

Visual Storytelling Visual Storytelling Dataset[7] (Huang et al., 2016) contains ∼81K unique images obtained from Flickr. Each album contains image sequences comprising 5 photos and 5 human-written stories, and each story usually comprises one sentence per image. The unique number of stories is ∼50K. About 67% of the images also have matching human-written captions. We used the story sentences and captions to create the noun tag dictionary.

G Comprehensive Results: VST

Table 15 reports automatic evaluation results on the VIST benchmark. We compare baseline storytelling systems with variants augmented using MMCOMET-based multimodal commonsense retrieval. Incorporating MMCOMET improves visual grounding (R-VG), coherence (R-C), and non-redundancy (R-NR). The best-performing configuration (LLaVA+MMCOMET^{S+I}) achieves the highest overall composite score (R+S+B+M+U = 272.0), substantially outperforming non-augmented vision-language baselines. Notably, improvements are most pronounced in grounding-sensitive metrics (R-VG) and semantic similarity measures (BLEURT, MoverScore), indicating that multimodal commonsense contributes additional contextual cues beyond raw visual features.

Table 15: Automatic evaluation results of visual storytelling. MMCOMET-enhanced models (bottom rows) outperform baselines in visual grounding, coherence, and overall story quality metrics, demonstrating the benefit of multimodal commonsense knowledge (for **Perplexity**, lower score is better.)

Model	R-VG	R-C	R-NR	RoViST (R)	SPICE (S)	BLEURT (B)	MoverScore (M)	UNION (U)	Perplexity	R+S+B+M+U	Story L.
SCO-VIST	70.3	72.3	90.5	77.7	**10.9**	34.9	56.0	84.0	14.9	263.5	52.3
VILT	70.2	76.3	94.2	80.2	9.3	34.3	52.1	83.3	12.1	259.2	50.1
LLaVa	72.3	78.1	96.1	82.2	9.5	34.1	53.9	**84.2**	10.5	263.9	51.9
+MMCOMET^{S}	73.3	78.4	95.8	82.5	9.9	35.2	54.2	82.6	**10.8**	264.4	50.8
+MMCOMET^{S+I}	**75.2**	**79.1**	**96.2**	**83.5**	10.2	**36.7**	**57.4**	**84.2**	11.2	**272.0**	53.4

[3]https://visualcommonsense.com/

[4]https://huggingface.co/datasets/HuggingFaceM4/VQAv2

[5]https://huggingface.co/datasets/nlphuji/flickr30k

[6]https://cure-lab.github.io/PnPInversion/

[7]https://visionandlanguage.net/VIST/dataset.html

H Ablation Studies

H.1 VCR Category and Ambiguity Analysis

Breakdown Categories Following VCR (Zellers et al., 2019), we categorise VCR questions into seven categories based on keyword patterns: Explanation (37.5%), Activity (14.7%), Temporal (4.9%), Mental (5.1%), Role (0.5%), Scene (9.9%), Hypothetical (4.7%). Questions not matching any keyword pattern are classified as Others (22.8%). As the authors note that the complete keyword patterns and code were not released, we manually inspected the questions and extended the keyword list for each category accordingly. Figure 3 shows representative examples of frequently retrieved phrases and images for each category. We excluded the Role category because it has a significantly lower percentage compared to other categories.

To further investigate when multimodal commonsense knowledge is most beneficial, we analyse model performance based on ambiguity, defined as $Ambiguity = 1 - (p(top_1) - p(top_2))$, where $p(top_1)$ and $p(top_2)$ denote the predicted probabilities assigned to the top-1 and top-2 answer candidates by the baseline model, respectively. A higher ambiguity score indicates that the model assigns similar probabilities to multiple candidates, reflecting samples where the correct answer is less distinguishable, or inherently harder samples.

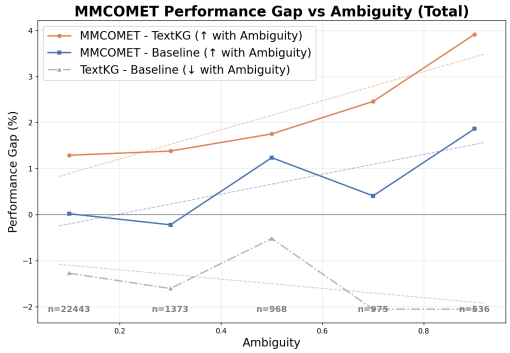


Figure 6: Ambiguity vs Performance Gap

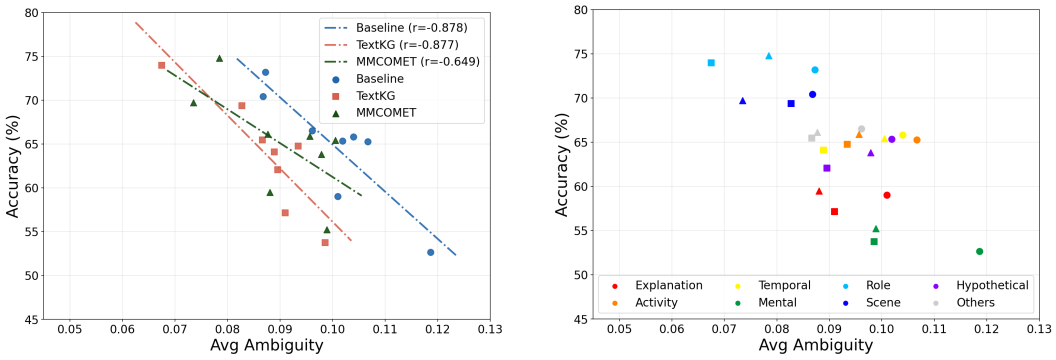


Figure 7: Ambiguity vs Accuracy by Method (Left) and Category (Right).

Ambiguity Analysis Figure 7 shows the correlation between average ambiguity and accuracy per category for each method. While Baseline and TextKG exhibit strong negative correlation ($r = -0.878$ and $r = -0.877$, respectively), MMCOMET shows a notably weaker correlation ($r = -0.649$), indicating that multimodal grounding partially mitigates the performance degradation associated with high ambiguity. Together, these results suggest that MMCOMET is particularly effective in challenging scenarios, providing more reliable commonsense augmentation precisely when the model is most uncertain. Also it reveals that across most categories including Explanation,

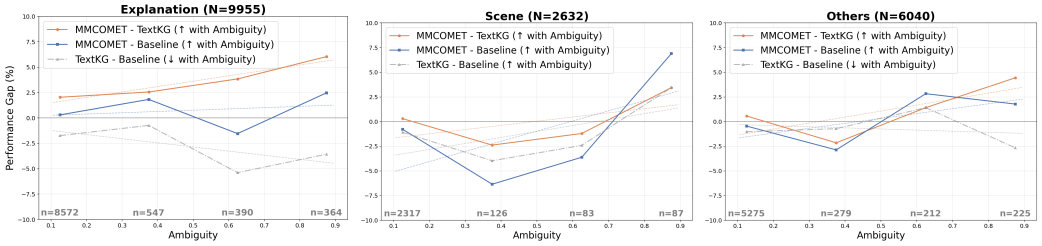


Figure 8: Ambiguity vs performance gap for representative categories.

Mental, and Activity, MMCOMET reduces average ambiguity while simultaneously improving accuracy compared to the baseline, suggesting that multimodal commonsense knowledge helps the model become more confident and more correct across diverse question types. In contrast, TextKG tends to reduce ambiguity but also degrades accuracy in many categories, indicating that TextKG often acts as noise.

Figure 6 shows how the performance gap between methods varies across ambiguity bins. As ambiguity increases, the gap between MMCOMET and TextKG grows consistently, reaching about +4% in the highest ambiguity bin, while the gap between MMCOMET and Baseline also increases monotonically. In contrast, the gap between TextKG and the baseline widens in the negative direction, indicating that TextKG becomes increasingly detrimental as the sample difficulty increases. This asymmetry suggests that visual grounding is especially critical for harder samples, where text-only signals fail to identify contextually relevant knowledge. We present the performance gap across representative categories in Figure 8. Across Explanation, Scene, and Others, the gap between MMCOMET and TextKG generally increases with ambiguity, consistent with the trend observed in the total analysis. While some intermediate bins show performance drops, we attribute this largely to the uneven sample distribution across bins rather than a systematic degradation.

H.2 Domain Coverage

Table 16: Domain contribution to retrieved phrases and leave-one-out performance impact, for VCR, VQA, and Image Captioning. P/S/E denote Physical, Social, Eventive.

Domain	Phrase Share (%)			Leave-one-out Performance						
	VCR	VQA	IC	VCR		VQA		Image Captioning		
				InternVL38B	Qwen8B	Gemma31B	Qwen8B	BLEU-4	CIDEr	ROUGE-L
P	20.71	40.66	40.87	**75.3**	**66.5**	65.65	69.13	0.169	0.357	0.443
S	56.13	46.96	53.56	74.9	66.4	66.05	69.68	0.173	**0.372**	**0.448**
E	54.04	35.06	32.54	75.1	66.3	66.1	68.75	**0.174**	0.362	0.447
All (Full)	-	-	-	**75.3**	66.4	**66.31**	**69.74**	0.17	**0.372**	0.446

We extend our analysis to quantify downstream gains based on joint domain coverage for VCR and VQA covering discriminative tasks, and for Image Captioning covering the generative task. The distribution of the retrieved phrases based on their associated domain/s is provided in Table 16. We note that each phrase may belong to one or more domain, resulting in overlapping distributions.

For VCR, excluding the physical domain showed no performance loss. The domain distribution of the retrieved phrases for each task (Table 16) reveals that the VCR dataset heavily relies on the Social and Eventive domains, with more than 50% of phrases. In comparison, the Physical domain accounts for only around 20%. This demonstrates the significance of MMCOMET's social and eventive dimensions for visual reasoning. The VQA task and dataset show a more evenly distributed domain distribution, resulting in better performance when all domains are combined. For Image Captioning, excluding the physical domain degrades performance most, despite the social domain matching more than 50% of phrases. This is because captions must ground in visible image components, regardless of depicted social aspects.

H.3 Alignment Strategy

Table 17 isolates each alignment method's benefits using Qwen3-8B. For VCR, removing similarity-matched images (focused on visually concrete concepts) hurts performance most. Conversely, Image Captioning benefits more from abstract social or eventive concepts via web-based retrieval, improving without caption-sourced images. These results highlight the varying requirements of different downstream tasks, which benefit from distinct visual grounding routes. MMCOMET's hybrid construction pipeline provides broad coverage, encompassing both concrete and abstract concepts, rather than guaranteeing a task-specific optimum.

Table 17: Alignment Strategy Ablation Study for VCR and Image Captioning downstream tasks.

Alignment		VCR	Image Captioning		
Similarity	Web Only	Accuracy	BLEU-4	CIDEr	ROUGE-L
o	x	64.4	0.169	0.371	0.445
x	o	63.1	**0.177**	**0.387**	**0.452**
o	o	**66.4**	0.17	0.372	0.446

H.4 Effect of Knowledge Source

Table 18: Performance comparison against VTKG-C for VCR, VQA, Image Captioning(CIDEr) downstream tasks.

Setup	VCR		VQA		IC	
	InternVL38B	Qwen8B	Qwen8B	Gemma31B	Qwen8B	Qwen32B
Baseline	73.8	65.3	68.60	64.72	0.232	0.151
TextKG	74.1	62.7	68.41	65.55	0.346	0.147
VTKG-C	73.8	64.1	**70.01**	64.85	0.323	0.124
MMCOMET	**75.3**	**66.4**	69.74	**66.31**	**0.372**	**0.157**

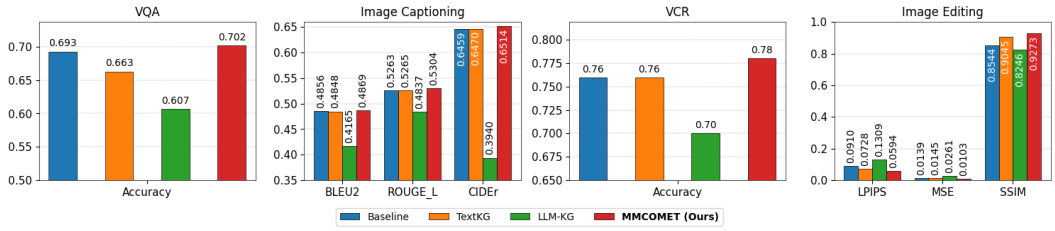


Figure 9: Incorporation of MMCOMET against baseline, TextKG, and LLM generated knowledge.

LLM-Generated Knowledge As discussed in Sec. 5.7, we compare MMCOMET against knowledge generated by prompting InternVL3-8B to produce commonsense phrases about objects, people, relationships, behaviors, and the contextual background of an input image, using 100 samples per task. Figure 9 shows that MMCOMET outperforms all other setups for VQA and VCR, where model reasoning benefits from the integrated physical, social, and eventive knowledge. LLM-generated knowledge performs worst across all tasks, indicating that the structured knowledge from the commonsense graphs better directs visual grounding than naive knowledge integration. Perception-dominated tasks such as IC and IE benefit less from commonsense knowledge integration overall; however, MMCOMET still achieves greater gains than text-only knowledge, demonstrating more robust knowledge.

VTKG-C We evaluate MMCOMET against VTKG-C, an existing multimodal commonsense knowledge graph limited to the physical domain, on three representative settings that jointly span the task space: VCR (reasoning-heavy), VQA (more general, perception-oriented), and Image Captioning (generative). Table 18 report results for VCR/VQA and Image Captioning, respectively. Injecting knowledge from VTKG-C does not benefit VCR, resulting in similar or even degraded performance compared to the baseline. On VQA, VTKG-C yields a marginal gain over MMCOMET for Qwen8B (70.01 vs. 69.74) but underperforms for Gemma31B, suggesting its physical-domain knowledge occasionally aligns with VQA's more perception-oriented questions but does not generalize across models. For Image Captioning, VTKG-C improves upon the baseline but falls short when compared to using the text-based TextKG method. Overall, MMCOMET outperforms VTKG-C on the majority of tasks and models, demonstrating the importance of extending commonsense knowledge beyond the physical domain to include social and eventive dimensions.

I EXTRINSIC QUALITATIVE ANALYSIS

To further illustrate the effect of using MMCOMET as auxiliary information in visual-language downstream tasks, we show a selection of representative examples in the following sections.

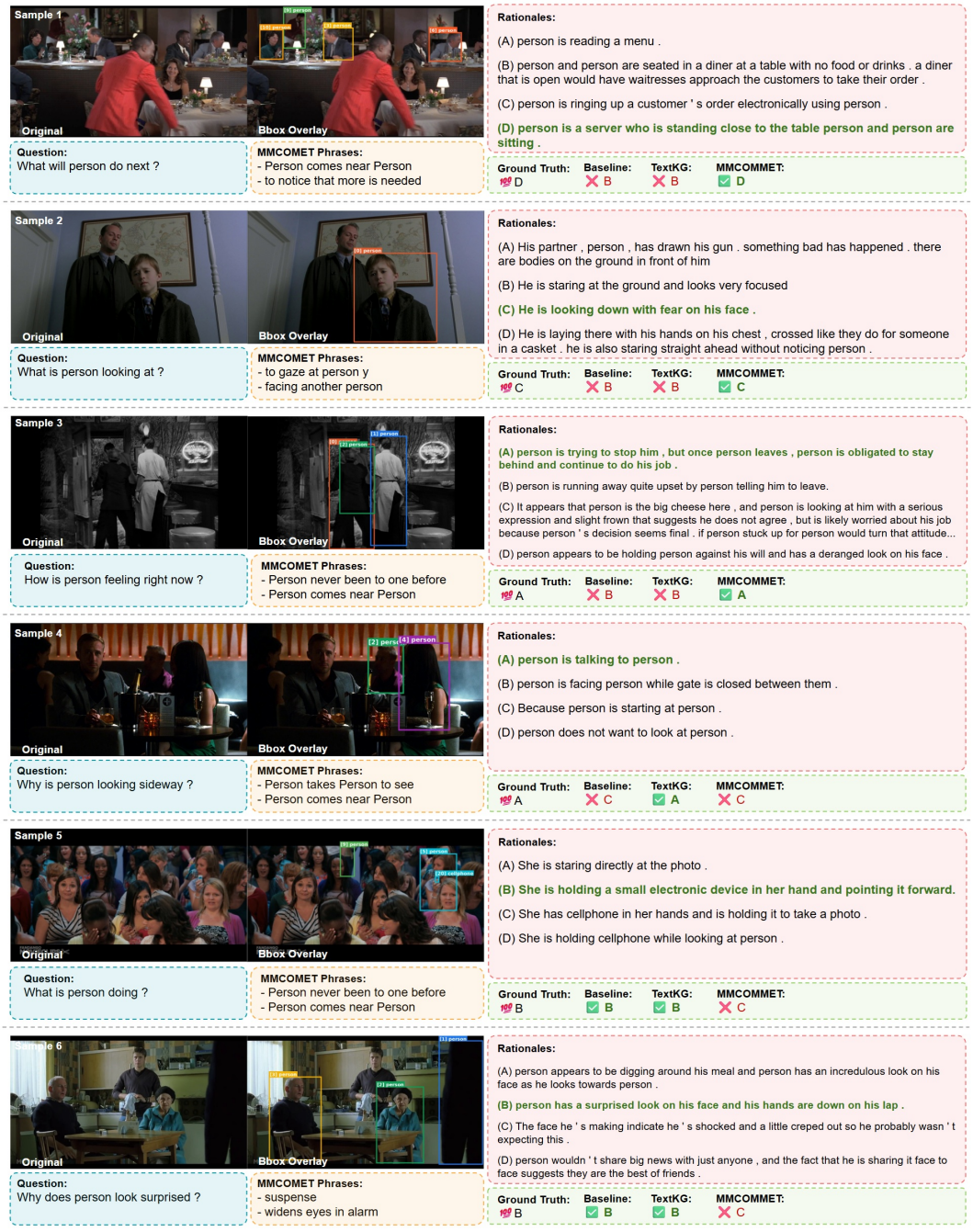


Figure 10: Qualitative samples for Visual Commonsense Reasoning.

I.1 VISUAL COMMONSENSE REASONING

Figure 10 shows representative examples of VCR tasks, where the upper three cases are success cases and the lower three cases are failure cases. Taken together, these examples reveal that base-

line models typically select rationales based on salient features. In contrast, MMCOMET output demonstrates richer commonsense grounding that can be grouped into two reasoning patterns. **(1) Perceptual state inference.** MMCOMET enables the model to reason about internal states that are not directly readable from superficial features. For instance, in sample 2, both the Baseline and TextKG fail to identify the correct rationale, selecting answers that reflect only surface-level scene description. MMCOMET, guided by retrieved phrases encoding perceptual attention of correctly infer based on the emotional quality underlying that gaze. **(2) Social and emotional context modeling.** Enhanced outputs incorporate implicit interpersonal dynamics that require reading beyond individual characters. In such cases, MMCOMET leverages phrases that capture proximity and relational roles to correctly identify the situation, discriminate among visually similar but socially distinct rationales, and arrive at the interpretation that best reflects the underlying situation.

Similarly, MMCOMET's failures tend to fall into two patterns. First, retrieved phrases are broadly relevant to the image yet semantically misapplied, for example, in sample 6, the phrase 'widens eyes in alarm' is successfully retrieved in a surprise scenario, but the model nonetheless selects an incorrect rationale centered on social inference, suggesting that retrieval relevance does not guarantee correct integration into downstream reasoning. Second, in cases where the question is based on fine-grained object-level or action-level details, models fail to connect meaningfully with the visual evidence in the image.

I.2 Visual Question Answering



Figure 11: Qualitative samples for Visual Question Answering.

Figure 11 presents representative VQA examples where both Baseline and TextKG produce incorrect answers while MMCOMET succeeds. These cases illustrate how commonsense retrieval supports correct inference.

In the middle example, the retrieved phrases'serve the ball' and 'He plays ball with them' do not directly address the question about color. However, by anchoring the model to the correct activity context, these phrases appear to facilitate more focused visual attention, enabling the model to localise the relevant region and correctly identify its color. This suggests that commonsense retrieval can also benefit the model by providing contextual scaffolding that sharpens the focus on question-relevant visual details.

In the right example, it highlights a complementary failure mode of the baseline. When asked what the traffic light says, the baseline responds with the visual feature only. However, both TextKG and MMCOMET correctly answer stop, suggesting access to commonsense knowledge about traffic light conventions. It seems that the retrieved commonsense phrases 'wait for green lights' and 'waits for light to turn green' help the model interpret the question at the appropriate level of abstraction, shifting from perceptual description to functional meaning.

I.3 Visual Story Telling

Figure 12 presents representative Visual storytelling examples. Baseline outputs tend to enumerate observed visual elements as isolated facts, producing narratives that are factually grounded but experientially flat. MMCOMET, in contrast, introduce richer narrative structure by embedding social and eventive context around the same visual observations.

The distinction is evident across all three examples. In the farm visit story, the baseline notes the presence of animals and a ride. At the same time, MMCOMET reconstructs the scene as a shared experience, situating individual observations within a coherent narrative of the outing. In the party



Baseline:
Today we visited a farm. We saw animals. We laughed together. The wind blew. We enjoyed a ride.

MMCOMET (LLaVa+MMCOMET^{S+I}):
Today we visited a farm where we spotted some pigs at the table. We saw some playful pigs in the pen. At the picnic table, my brother and I shared laughs. The windmills nearby added a gentle breeze to the air. We enjoyed ourselves with laughter and excitement.



Baseline:
I attended a party recently. Food was served. Neighbors were guests. Animals were present.
A cake concluded the event.

MMCOMET (LLaVa+MMCOMET^{S+I}):
I went to the party last night. We had a great time with many delicious dishes prepared by our friends. We all enjoyed ourselves, chatting and laughing. The dogs were so excited they couldn't stop wagging their tails. We cut the cake at the end, ending the evening on a sweet note.



Baseline:
The venue is prepared for the celebration. The groom receives guidance from his friend. A ring is placed carefully in its box. The bride and her attendant take a photo. Guests savor the festivities.

MMCOMET (LLaVa+MMCOMET^{S+I}):
The room is ready for the wedding reception, adorned with twinkling lights.
The best man offers the groom some guidance on his upcoming journey as a new husband.
An engagement ring gleams inside its elegant box, waiting to be presented.
The bride and maid of honor laugh as they pose for a candid portrait, capturing their joyful moment.
As guests savor delicious food, everyone enjoys the celebration, filled with love and laughter.

Figure 12: Qualitative samples for Visual Storytelling.

example, the baseline lists food, guests, and a concluding cake. In contrast, MMCOMET elaborates on the social atmosphere, capturing collective enjoyment through details such as chatting, laughing, and excited dogs. Most notably, in the wedding story, MMCOMET transforms each visual panel into a narratively connected moment that is not directly observable from the images alone.

Taken together, these examples suggest that commonsense retrieval enables the model to move beyond visual enumeration toward socially and eventively grounded storytelling, producing narratives that better reflect how people experience and recall shared events.

I.4 IMAGE EDITING

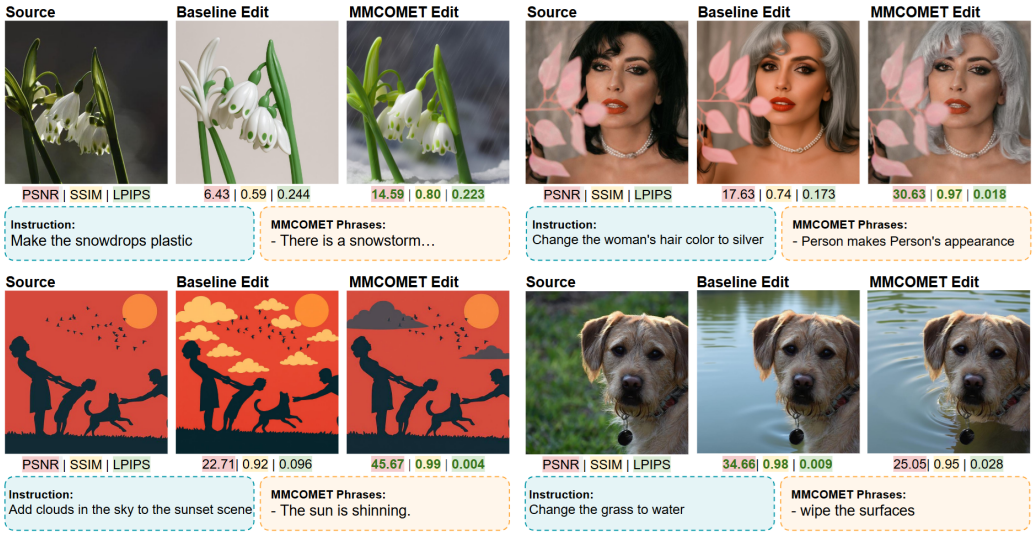


Figure 13: Qualitative samples for Image Editing.

Figure 13 presents representative image editing examples. Across all cases, MMCOMET consistently achieves better source preservation than the baseline, as reflected in higher PSNR and SSIM scores, while applying targeted modifications that align with the instruction's intent (e.g., correctly localising appearance changes to the relevant region while leaving surrounding content intact in the upper-right sample). This pattern suggests that commonsense retrieval encourages conservative editing, prioritising structural fidelity over aggressive transformation. In the case of 'Change the grass to water', the retrieved phrase 'wipe the surfaces' provides only a generic surface-level association, failing to capture the material and textural transformation required. Consequently, MMCOMET falls short of edit fidelity in the automatic metric, yet still introduces subtle surface undulations that are more consistent with the appearance of water than the baseline output, suggesting that, even in failure cases, the model attends to semantically relevant visual properties rather than defaulting to arbitrary modifications.

I.5 IMAGE CAPTIONING



Figure 14: Qualitative samples for Image Captioning.

Figure 14 presents representative image captioning examples. Baseline-generated captions tend to over-describe objects that are not the main focus of the image, such as "black jacket and shorts" in the top-right image and "near a red vehicle" in the bottom-right image. In contrast, MMCOMET's captions consistently move beyond surface-level visual description toward more contextually grounded characterisations. Rather than enumerating appearance and spatial arrangement, MMCOMET outputs capture the underlying intent, role, or interactional quality of the depicted scene.

This pattern is evident in two complementary ways. In the marathon example, the baseline describes the runner's appearance and physical context in detail, yet fails to identify the event itself. MMCOMET, guided by retrieved phrases, correctly situates the action within a specific event schema, producing a caption that reflects both appearance and context. In the baby example, both the baseline and TextKG describe the scene as a static holding relationship, whereas MMCOMET reframes the interaction as holding and engaging with a baby, informed by retrieved phrases that encode interpersonal engagement with an infant.

J Intrinsic Qualitative Analysis

J.1 Alignment Quality Analysis

Table 2 shows the result of human evaluation on image-phrase pairs, evaluating alignment quality.

Result by retrieval source Web-based retrieval (WSE) consistently outperforms similarity matching over open-source caption datasets (OID). Aggregated across all relations, WSE achieves 91.5% (FM+PM) and 79.0% (FM), compared to 82.4% and 61.2% for OID. This difference reflects the advantage of query-sensitive search for abstract and compositional phrases, while caption datasets often lack explicit coverage of eventive or social expressions. Approximately 72% of phrases in MMCOMET are grounded via web retrieval, underscoring the importance of the hybrid routing strategy.

Result by relation types Alignment quality also varies across relation types. Physical relations achieve the highest grounding rates (97.8% FM+PM), followed by eventive (83.3%) and social (84.3%) relations. This trend aligns with phrase concreteness: tangible entities are easier to visually ground, whereas abstract emotional or intentional states exhibit greater ambiguity. Despite this inherent difficulty, even the most challenging relation types maintain alignment above 77% (FM+PM).

J.2 Fully vs. Partially Matched Images

Figure 15 shows examples of what we consider a full versus partial match for an image. An image is a full match if all the actions and concepts in the commonsense phrase are presented in the image. On the other hand, it is a partial match if some aspect is missing but the overall theme of the image still matches the commonsense phrase and will therefore still be useful if applied to the downstream task. In the first ('*PersonX gets in the car*') and third example ('*PersonX eats sandwiches for lunch*'), the left image is considered a full match as the images clearly show the action of a person entering a car and eating a sandwich respectively. Conversely, the images on the right are considered partial matches as for the first example, the image on the right just shows a person in the car but the action of 'getting into the car' is missing. Likewise, for the third example, the image on the right simply depicts sandwiches but the action of 'eating' is absent. For the second ('*PersonX gives toys for Christmas'*) and fourth example ('*PersonX loves to garden*'), an important noun or concept is missing from the partially matched images. Specifically, for the second example, although the action of 'giving' is present in the image on the right, the tangible noun 'toys' is not visually captured whereas the left image of a child holding a toy bear heavily implies she had received it from another person. Similarly, the image on the right in the fourth example is missing the important entity 'person' whereas the left image is a full match as it clearly depicts a person happy in his garden.



Figure 15: Examples of a fully matched image (left) and partially matched image (right) for each 4 different commonsense phrases.

J.3 Head-Relation-Tail Image Phrase Matching

Figure 16 shows examples of several relations, each with 3 matching image examples for each of the head and tail are shown. For instance, connecting "*gift*" and "*please person*" through images depicting both the physical form of gifts and the emotional responses they elicit. Eventive relations such as *isAfter* and *HasSubEvent* demonstrate temporal and causal consistency, where sequences like "*PersonX quickly made a sandwich*" and "*PersonX woke up later for work*" visually manifest cause-and-effect coherence.



Figure 16: Examples of retrieved images corresponding to selected head and tail phrases across multiple relation types. Blue, orange and green arrows indicate physical, eventive and social relations, respectively. Images with red borders indicate images obtained from open-source datasets, while images with blue borders indicate images retrieved through web searches.

K ADAPTIVE KNOWLEDGE UTILITY ROUTER

To further address the heterogeneous reasoning requirements of diverse downstream vision-language tasks, we propose an adaptive knowledge utility router that evaluates the usefulness of different knowledge sources or routes. Instead of relying on hard, one-hot classification, which treats route utility as a binary switch, discarding complementary information from overlapping paths, the adaptive knowledge router models the utility of each route independently using a soft utility score. This multi-route formulation preserves overlapping successful routes, accounts for potentially harmful interventions, and allows the model to default to a no-knowledge prediction when additional knowledge yields no benefit. The adaptive knowledge routing framework has three stages (Algorithm 1 and 2): 1) oracle-derived utility construction, 2) utility-router optimization, and 3) routed inference.

K.1 ROUTER COMPONENT CONSTRUCTION

During utility component construction, a frozen VLM model extracts all baseline and knowledge-augmented components for each sample. To exhaust the commonsense knowledge coverage of MMCOMET, we utilitze its individual domains and their combinations as 7 distinct routes $r \in \mathcal{R}_K$: Physical Only (P), Social Only (S), Eventive Only (E), P+S, P+E, S+E, and All. The baseline components includes the base prediction $\hat{y}_i^0$, the relevant token logits p_i^0, and the base confidence profile s_i, consisting of entropy, maximum predicted probability, and confidence margin. For each sample and knowledge route, we extract the augmented prediction $\hat{y}_i^r$, soft utility score u_i^r, and encoded state z_i^r. The soft utility score represents the net gain in the log-probability gap between the ground truth and the strongest incorrect answer across the base and the knowledge-augmented models. An independent utility encoder using a Qwen3- VL-Embedding-8B backbone constructs the encoded state that serves as a a route-specific representation derived from the input image, question, retrieved MMCOMET phrases, base confidence profile, and route information. The soft utility score and the encoded state serve as the training inputs and targets for training the knowledge utility router.

K.2 ROUTER OPTIMIZATION AND INFERENCE

The knowledge utility router is a lightweight, Transformer-based model that jointly predicts the utility scores for all routes $u_i^{\mathcal{R}_K}$ for each sample within a minibatch. To mitigate the impact of outliers and extreme margin values during training, we constrain all utility scores using a hyperbolic tangent function parameterized by a temperature hyperparameter $\mathcal{T} \in \{2, 5, 10\}$ during training. Optimization is guided by a joint objective consisting of Huber loss for regression and a within-sample pairwise margin-ranking loss to effectively identify optimal routes. During inference, when the maximum predicted utility score falls below a threshold $\tau \in \{-0.1, -0.01, 0, 0.01, 0.1\}$, the router bypasses external knowledge and returns the base answer, indicating that no knowledge is beneficial. Otherwise, knowledge from the highest-utility route is used to generate the final prediction.

Table 19: Relevant Router Hyperparameters

Hyperparameter	Value
Utility Encoder	
hidden_dim	512
output_dim	512
Adaptive Router	
n_head	8
num_layers	3
dim_feedforward	1024
dropout	0.1
learning_rate	1e-05
epochs	100
early_stop	5

K.3 ADAPTIVE ROUTING SETUP

We evaluate our proposed adaptive router through a 5-fold cross-validation on MMStar[8] (N=500) to ensure robustness. MMStar is a comprehensive VQA dataset that features six distinct question types with different knowledge requirements, including Coarse Perception (CP), Fine-grained Perception (FP), Instance Reasoning (IR), Logical Reasoning (LR), Math (MA), and Science and Technology (ST). We use different model families: Qwen3-VL-8B-Instruct, Qwen3-VL-32B-Instruct, LLaVa1.6-34B, and Mistral3.1-24B. For MMCOMET knowledge retrieval during each knowledge-augmented route, we maintain the same hyperparameters as outlined in Appendix F. For the utility encoder, we use Qwen3-VL-Embedding-8B as the backbone to extract encodings for the image,

[8]https://mmstar-benchmark.github.io/

Algorithm 1 Adaptive Knowledge Utility Router Training

Require: Training set $\mathcal{D}$, frozen VLM M, MMCOMET $\mathcal{G}$, knowledge routes $\mathcal{R}_K$, utility router U_θ, scaling temperature T
Ensure: Trained utility router U_θ

Oracle-derived utility construction

for all $(I_i, Q_i, y_i) \in \mathcal{D}$ **do**
  $(\hat{y}_i^0, p_i^0) \leftarrow M(I_i, Q_i)$
  $s_i \leftarrow [H(p_i^0), \max(p_i^0), p_{i,(1)}^0 - p_{i,(2)}^0]$
  for all $r \in \mathcal{R}_K$ **do**
    $K_i^r \leftarrow \textsc{Retrieve}(\mathcal{G}, I_i, Q_i, r, B)$
    $(\hat{y}_i^r, p_i^r) \leftarrow M(I_i, Q_i, K_i^r)$
    $u_i^r \leftarrow \left[\log p_i^r(y_i) - \max_{y_i' \neq y_i} \log p_i^r(y_i')\right] - \left[\log p_i^0(y_i) - \max_{y_i' \neq y_i} \log p_i^0(y_i')\right]$
    $z_i^r \leftarrow \textsc{Encode}(I_i, Q_i, K_i^r, s_i, r)$
  end for
end for

Utility-router optimization

repeat
  for all $(z_i^{\mathcal{R}_K}, u_i^{\mathcal{R}_K})$ in a minibatch **do**
    $\hat{u}_i^{\mathcal{R}_K} \leftarrow U_\theta(z_i^{\mathcal{R}_K})$
  end for
  ${u'}_i^{\mathcal{R}_K} \leftarrow \textsc{tanh}(u_i^{\mathcal{R}_K}/T)$
  $\mathcal{L}_{\text{huber}} \leftarrow \textsc{HuberLoss}({u'}_i^{\mathcal{R}_K}, \hat{u}_i^{\mathcal{R}_K})$
  $\mathcal{L}_{\text{rank}} \leftarrow \textsc{MarginRankingLoss}({u'}_i^{\mathcal{R}_K}, \hat{u}_i^{\mathcal{R}_K})$
  Update θ using $\mathcal{L}_{\text{huber}} + \lambda_{\text{rank}}\mathcal{L}_{\text{rank}}$
until convergence
return U_θ

Algorithm 2 Adaptive Knowledge Utility Router Inference

Require: Input (I, Q), frozen VLM M, MMCOMET $\mathcal{G}$, router U_θ, knowledge routes $\mathcal{R}_K$, utility threshold τ
Ensure: Prediction $\hat{y}$

$(\hat{y}^0, p^0) \leftarrow M(I, Q)$
$s \leftarrow [H(p^0), \max(p^0), p_{(1)}^0 - p_{(2)}^0]$
for all $r \in \mathcal{R}_K$ **do**
  $K^r \leftarrow \textsc{Retrieve}(\mathcal{G}, I, Q, r, B)$
  $z^r \leftarrow \textsc{Encode}(I, Q, K^r, s, r)$
end for
$\hat{u}^{\mathcal{R}_K} \leftarrow U_\theta(z^{\mathcal{R}_K})$
$r^* \leftarrow \arg\max_{r \in \mathcal{R}_K} \hat{u}^r$
if $\hat{u}^{r^*} \leq \tau$ **then**
  return $\hat{y}^0$ ▷ No knowledge required
else
  $\hat{y} \leftarrow M(I, Q, K^{r^*})$
  return $\hat{y}$
end if

question, knowledge phrases, route information, and confidence profile. Each component is individually encoded, then concatenated, and fused via linear layers. Table 19 outlines the relevant hyperparameters. We use accuracy as our primary performance metric.

K.4 Adaptive Routing Results

Table 20 shows the overall and per-category accuracy between the no-knowledge baseline, full MMCOMET integration, and MMCOMET with the proposed adaptive, dynamic router. On this dataset, the full MMCOMET framework degrades overall baseline performance, primarily driven by the heterogeneity of question types. Despite this, MMCOMET noticeably outperforms the baseline in Logical Reasoning (LR) question types for larger models Qwen3-32B (+0.0441) and LLaVa1.6-34B (+0.0735), demonstrating that MMCOMET makes retrieved commonsense knowledge more

Table 20: MMStar categorical performance comparison between baseline, full MMCOMET, and MMCOMET with the adaptive router. CP: Coarse Perception; FP: Fine-grained Perception; IR: Instance Reasoning; LR: Logical Reasoning; MA: Math; ST: Science and Technology.

Model	Setup	Overall	CP	FP	IR	LR	MA	ST
Qwen3-32B	Baseline	0.7160	0.7614	**0.7412**	**0.8161**	0.6912	0.7500	**0.5341**
	MMCOMET (Full)	0.7000	0.7500	0.7176	0.7931	0.7206	0.7024	0.5227
	MMCOMET + Router	**0.7260**	**0.7841**	**0.7412**	0.8046	**0.7353**	**0.7619**	**0.5341**
LLaVa1.6-34B	Baseline	0.5060	**0.6364**	**0.4588**	0.6552	0.4265	**0.4881**	**0.3523**
	MMCOMET (Full)	0.4720	0.6023	0.3882	0.6437	0.4706	0.4405	0.2841
	MMCOMET + Router	**0.5080**	**0.6364**	0.4000	**0.7011**	**0.5000**	0.4762	0.3295
Mistral	Baseline	**0.6020**	0.6818	0.5176	**0.7241**	0.5735	0.6071	**0.5000**
	MMCOMET (Full)	0.5840	0.6932	0.5294	**0.7241**	0.5441	0.5952	0.4091
	MMCOMET + Router	**0.6020**	**0.7045**	**0.5412**	0.6897	**0.5882**	**0.6310**	0.4545
Qwen3-8B	Baseline	0.6460	**0.7386**	**0.6824**	0.6897	0.6176	0.6548	0.4886
	MMCOMET (Full)	0.6300	0.7045	0.6118	**0.7471**	0.6029	0.6548	0.4545
	MMCOMET + Router	**0.6520**	0.7159	0.6706	0.6897	**0.6324**	**0.6905**	**0.5114**

reliable, particularly for reasoning-heavy downstream tasks. Smaller models struggle more with MMCOMET alone, possibly due to their limited reasoning capabilities.

We note that performance varies across different model-category combinations, exhibiting differential sensitivities to the visually grounded commonsense knowledge from MMCOMET, highlighting the inherent heterogeneity of reasoning requirements in the dataset and the varying strengths of specific model types. Our adaptive routing mechanism is designed to exploit this heterogeneity by conditionally activating knowledge paths for each input based on predicted route utility and confidence. This dynamic routing not only selects the optimal path but also further mitigates extreme performance degradation from the rigid application of MMCOMET, resulting higher overall accuracies and mitigated effects from categories that do not benefit as much from commonsense knowledge, such as heavy perception-based tasks. This underscores the importance of dynamically bypassing knowledge paths when augmentation proves detrimental rather than beneficial to the downstream task.

L PROMPT TEMPLATES

Tables 21-31 provide all the prompt templates for Visual Commonsense Reasoning, Visual Question Answering, Image Editing, and Image Captioning.

Table 21: Baseline Prompt Template (VCR)

Baseline Prompt Template (VCR)

```
You are given an image, a question, and four candidate rationales.

Choose the rationale that is best supported by the image and that most plausibly
explains the situation in the question.
When deciding:
- Prioritise visible evidence in the image.
- Prefer the choice that best matches the people, objects, actions, and scene context.
- Reject choices that are too generic, unsupported, or contradicted by the image.
- Do not choose a rationale just because it sounds generally plausible.

{fewshot}

Question:
{question}
Candidate rationales:
(A) {r0}
(B) {r1}
(C) {r2}
(D) {r3}

Think silently about which option is most consistent with the image, then answer with
the letter only.
Answer:
```

Table 22: TextKG Prompt Template (VCR)

TextKG Prompt Template (VCR)

```
You are given an image, a question, four candidate rationales, and retrieved
commonsense knowledge.

Your goal is to choose the rationale that is best supported by the image and most
plausibly explains the situation. Use the retrieved knowledge only as weak background
knowledge. If the knowledge is irrelevant, vague, or not clearly supported by the
image, ignore it. Always prioritise the image and the candidate rationales over the
retrieved knowledge.

Retrieved commonsense knowledge:
{kg_list}
When deciding:
- First use the image to identify the most likely situation.
- Then compare all four rationales carefully.
- Prefer the rationale that is most specific to the visible scene.
- Do not choose a rationale only because it matches the retrieved knowledge.

{fewshot}

Question:
{question}
Candidate rationales:
(A) {r0}
(B) {r1}
```

```
(C) {r2}
(D) {r3}

Think silently about which option is most consistent with the image, then answer with
the letter only.
Answer:
```

Table 23: MMCOMET Prompt Template (VCR)

MMCOMET Prompt Template (VCR)

```
You are given an image, a question, four candidate rationales, and retrieved
multimodal commonsense knowledge from visually similar situations.

Choose the rationale that is best supported by the image and most plausibly explains
the situation.
The retrieved multimodal knowledge may help only when it matches the visible scene.
If it is generic, weakly related, or conflicts with the image, ignore it. Always
prioritise the image over the retrieved knowledge.

Retrieved multimodal commonsense knowledge:
{kg_list}

When deciding:
- Use the image as the main evidence.
- Compare all four rationales, not just the first plausible one.
- Prefer the rationale that best explains the specific scene, actions, and
relationships in the image.
- Avoid generic commonsense answers that are not clearly grounded in the image.

{fewshot}

Question:
{question}

Candidate rationales:
(A) {r0}
(B) {r1}
(C) {r2}
(D) {r3}

Think silently about which option is most consistent with the image, then answer with
the letter only.
Answer:
```

Table 24: Baseline Prompt Template (VQA)

Baseline Prompt Template (VQA)

```
You are given an image and a question.

Your task is to answer the question based only on the visual content.

{fewshot}

Question:
{question}

Answer with a short, concise response (1-3 words).
Answer:
```

Table 25: TextKG Prompt Template (VQA)

TextKG Prompt Template (VQA)

```
You are given an image, a question, and retrieved commonsense knowledge.

Your task is to answer the question based primarily on the image.

Use the retrieved knowledge only as weak background information. If it is irrelevant,
vague, or not supported by the image, ignore it.

Retrieved commonsense knowledge:
{kg_list}

When answering:
- First identify the relevant visual evidence in the image.
- Then consider whether the knowledge is helpful.
- Do not answer based only on the knowledge.
- Prefer answers that are directly grounded in the image.

{fewshot}

Question:
{question}

Answer with a short, concise response (1-3 words).
Answer:
```

Table 26: MMCOMET Prompt Template (VQA)

MMCOMET Prompt Template (VQA)

```
You are given an image, a question, and retrieved multimodal commonsense knowledge
from visually similar situations.

Your task is to answer the question based primarily on the image.

The retrieved multimodal knowledge may help only when it matches the visible scene.
If it is generic, weakly related, or conflicts with the image, ignore it.

Retrieved multimodal commonsense knowledge:
{kg_list}

When answering:
- Use the image as the main source of evidence.
- Consider the knowledge only if it helps explain the specific situation in the image.
- Avoid generic answers that are not grounded in the image.
- Prefer answers that reflect the actual scene, objects, and interactions.

{fewshot}

Question:
{question}

Answer with a short, concise response (1-3 words).
Answer:
```

Table 27: Baseline Prompt Template (Image Editing)

Baseline Prompt Template (Image Editing)

```
Background:
You are an advanced image editing assistant designed to edit images based on given
instructions.
Your task is to apply the most appropriate visual modification to the input image
based on the editing instruction provided.

Input Information:
Image: [input image]
Editing Instruction: {instruction}
```

Table 28: MMCOMET Prompt Template (Image Editing)

MMCOMET Prompt Template (Image Editing)

```
Background:
You are an advanced image editing assistant designed to edit images based on given
instructions.
Your task is to apply the most appropriate visual modification to the input image
based on the editing instruction and the provided commonsense knowledge.
Explicit commonsense knowledge consists of statements retrieved from a commonsense
knowledge source, representing the beginning (head) and the outcome (tail) of a
commonsense event.
Each statement is retrieved based on its relevance to both the input image and the
editing instruction.
Your objective is to integrate the commonsense knowledge with the editing instruction
to apply the most appropriate and contextually grounded visual modification.

Input Information:
Image: [input image]

Explicit Commonsense Knowledge:
Using the following knowledge graph in head-tail format: {head} → {tail}

Editing Instruction: {instruction}
```

Table 29: Baseline Prompt Template (Image Captioning)

Baseline Prompt Template (Image Captioning)

```
Write a natural one-sentence caption for the image.

{fewshot}

Caption:
```

Table 30: TextKG Prompt Template (Image Captioning)

TextKG Prompt Template (Image Captioning)

```
You are given an image and some retrieved commonsense hints.
Write a natural one-sentence caption for the image.
Use the image as the primary source of information.
Use the commonsense hints only if they are clearly supported by the image.
Do not include details that are uncertain, speculative, or not visible.

{fewshot}

Retrieved commonsense hints:
{kg_list}

Caption:
```

Table 31: MMCOMET Prompt Template (Image Captioning)

MMCOMET Prompt Template (Image Captioning)

```
You are given an image and some retrieved multimodal commonsense hints.
Write a natural one-sentence caption for the image.
Use the image as the primary source of information.
Use the commonsense hints only if they are clearly supported by the image.
Do not include details that are uncertain, speculative, or not visible.

{fewshot}

Retrieved multimodal commonsense hints:
{kg_list}

Caption:
```