CoVCR: Bridging Visual Narrative Gaps via Context Generation for Robust Commonsense Reasoning

Xinyu Li, Shiliang Sun*

Supplementary Material

The supplementary material includes: 1) Additional descriptions of methods and implementation details, 2) Additional dataset and annotation details, 3) Additional results, 4) Additional qualitative examples. 5) Limitations and future work.

A.1. Additional Descriptions of Methods and Implementation Details

A.1.1. Dynamic Gating Network

To dynamically control the contribution of the external and internal contextual features, we introduce a scalar gating mechanism. Given the visual embeddings $\mathbf{F}_v \in \mathbb{R}^{k\times d_q}$ and the textual embeddings $\mathbf{F}_t \in \mathbb{R}^{L_t\times d_q}$, we first obtain the multimodal representation through a lightweight fusion layer:

$$\mathbf{z} = \phi(\mathbf{W}_v\mathbf{F}_v + \mathbf{W}_t\mathbf{F}_t + \mathbf{b}), \quad \mathbf{z} \in \mathbb{R}^{d_z}, \tag{1}$$

where $\mathbf{W}_v \in \mathbb{R}^{d_z\times k}$ and $\mathbf{W}_t \in \mathbb{R}^{d_z\times L_t}$ are learnable projection matrices, $\mathbf{b}$ is a bias term, and $\phi(\cdot)$ denotes a nonlinear activation function.

We then compute the scalar gate value via a softmax activation:

$$\lambda = \sigma\left(\mathbf{w}_g^\top \mathbf{z} + b_g\right), \quad \lambda \in (0,1), \tag{2}$$

where $\mathbf{w}_g \in \mathbb{R}^{d_z}$ and b_g are learnable parameters.

Finally, the gated contextual representation is obtained by interpolating between the internal context feature $\mathbf{Y}_I$ and the external context feature $\mathbf{Y}_E$:

$$\mathbf{Y} = (1-\lambda)\,\mathbf{Y}_I + \lambda\,\mathbf{Y}_E. \tag{3}$$

This gating mechanism enables the model to adaptively balance the reliance on internal contextual cues and external knowledge at each block.

A.1.2. Implementation Details

In both the visual context generation stage and the visual reasoning fine-tuning stage, we use a pretrained CLIP [8] ViT-L/336px as the visual encoder, with its text encoder serving as the context encoder. We adopt Qwen2-7B [9] and LLaMA3-8B [2] as the backbone language models. All experiments are implemented in PyTorch and trained on a single NVIDIA H20 GPU (80GB).

A.2. Additional Dataset and Annotation Details

A.2.1. Dataset Details for VCG Learning

The VisualCOMET dataset [7] is a large-scale resource for visual commonsense reasoning, providing structured commonsense inferences grounded to person-centric images. For each image, VisualCOMET annotates a Visual Commonsense Graph describing plausible past, present, and future events associated with specific people via co-reference tags (e.g., "Person1").

Each graph consists of three main types of textual inferences: **Events Before:** plausible events that happened prior to the depicted moment. **Events at Present:** descriptions of currently ongoing activities in the scene. **Events After:** plausible future events that may occur next.

In this work, we leverage VisualCOMET dataset specifically for visual context generation (VCG) learning:

Events at present correspond to internal context, capturing what is directly happening within the scene. In contrast, events before and after are treated as external context, providing temporal cues that help fill in missing or implicit cues of the visual narrative.

During VCG learning, the model is pretrained to generate these internal and external contextual descriptions directly from image inputs. This enables CoVCR to learn how to enrich multimodal representations by predicting cognitive-level scene context without requiring additional textual prompts, forming the foundation of the context-infused reasoning used in the main framework.

A.2.2. Additional Details on Visual-Context Sufficiency Annotation

To support our claim that CoVCR's improvements on VCR are due to handling insufficient visual narratives, rather than architectural bias or dataset artifacts, we provide further details on the human-annotated visual-context sufficiency evaluation.

*Corresponding author.

Score	Level	Description	Interpretation
1	**Fully Visually Sufficient (Easy)**	The image alone provides clear and complete information to answer the question correctly. The reasoning is straightforward and directly grounded in visible content.	**Answer can be derived purely from visual evidence**; no additional commonsense reasoning is needed.
2	**Largely Visually Sufficient**	The image provides enough visual cues to reasonably infer the correct answer, though minor commonsense reasoning may help confirm it.	**Mostly visual-based reasoning**, with light contextual inference.
3	**Moderately Visually Sufficient**	Both visual information and external commonsense knowledge are required. The image supports the question partially, but some reasoning beyond the visible scene is necessary.	**Balanced visual and commonsense reasoning.**
4	**Largely Visually Insufficient**	The image gives limited clues; the answer relies heavily on external knowledge or situational inference beyond what is shown.	**Minimal visual grounding**, mostly commonsense-dependent.
5	**Fully Visually Insufficient (Difficult)**	The image provides almost no relevant information to answer the question. The correct answer depends entirely on non-visual reasoning or assumptions about the unseen context.	**No reliable answer can be inferred from the image itself.**

Table 1. Five-level scale of visual narrative sufficiency for VCR dataset annotation.

Task Prompt:

Given an image, generate contextual descriptions to support reasoning.

Please provide:
- **Internal Context:** describe what is directly happening within the scene.
- **External Context:** describe plausible events before and after this moment that help explain or extend the scene.

Ensure the descriptions are concise, plausible, and grounded in the image.

Return your answer strictly in the following JSON format:
```
{
 "Internal": ...
 "External": ...
}
```

Figure 1. Prompt Design for Visual Context Generation Learning.

As described in Table 1, we categorize each VCR sample into one of five levels based on the degree to which the question can be answered using the visible content alone. This scale ranges from Level 1 (Fully Visually Sufficient), where the answer is directly observable from the image, to Level 5 (Fully Visually Insufficient), where the image offers almost no information relevant to answering the question and the reasoning must rely almost entirely on external commonsense knowledge or contextual inference.

For supplementary evaluation, five human annotators independently assessed 300 randomly sampled VCR items along the five-level scale. The annotators were instructed to judge how much of the required reasoning was grounded in the visible scene versus how much depended on commonsense or external context. Cases where the image contains explicit, unambiguous evidence were labeled as Levels 1–2, whereas cases requiring substantial inference about unseen events or commonsense were labeled Levels 4–5. Disagreements were resolved via majority voting to ensure consistency.

A.3. Additional Results

A.3.1. Comparison with State-of-the-art Methods

As shown in Table 2, the proposed CoVCR framework achieves competitive performance on the VCR benchmark, although its results are slightly lower than those of ViP-LLaVA and GPT4RoI. We attribute this performance gap to differences in pre-training scale and model adaptation strategies.

Both ViP-LLaVA and GPT4RoI benefit from large-scale multimodal pre-training and full-parameter fine-tuning, which enable stronger multimodal reasoning, alignment, and generalization capabilities. In contrast, due to computational resource constraints, our approach conducts context generation–oriented pre-training solely on the Visual-COMET dataset rather than on broader multimodal corpora. Moreover, for language model adaptation, we adopt a parameter-efficient LoRA fine-tuning strategy instead of

Task Prompt:

You are an evaluator for assessing the quality of context generated by a visual context generation (VCG) module. The generated context is used to assist reasoning in visual commonsense reasoning (VCR) tasks, where different samples require different degrees of external commonsense or contextual knowledge. The difficulty of each sample ranges from Level 1 to Level 5: Level 1 means the question can be answered directly based on image content without additional context, while Level 5 means the visual information alone is insufficient and rich external context is needed.

Given an input image, the difficulty level, a question, a ground-truth answer, and the VCG-generated context, your task is to evaluate the quality of the generated context. You should assess the context along the following five dimensions. Each score must be an integer from 1 to 5, where 1 = very poor and 5 = excellent:

Correctness: Whether the context is factually accurate and free from hallucinations or incorrect descriptions.
Informativeness: Whether the context provides additional useful information beyond the visible image.
Relevance: Whether the context is closely related to the image, the question, and the required reasoning.
Helpfulness: Whether the context meaningfully helps answer the question more accurately.
Overall: The overall utility and quality of the context for assisting reasoning.

Return your answer strictly in the following JSON format:

```
{
 "Correctness": x,
 "Informativeness": x,
 "Relevance": x,
 "Helpfulness": x,
 "Overall": x,
 "Explanation": "A brief explanation (3–6 sentences) justifying your scores."
}
```

Figure 2. Example GPT-4V prompt for evaluating VCG-generated context in visual commonsense reasoning, assessing correctness, informativeness, relevance, helpfulness, and overall quality using a structured JSON output.

full-parameter optimization, which substantially reduces computational cost but also limits performance gains.

Despite this lightweight training paradigm, CoVCR still surpasses multiple strong baselines and demonstrates clear effectiveness in enhancing contextual reasoning through generated internal and external cues. These results suggest that context enhancement is an orthogonal and complementary direction to model scaling, offering tangible benefits to existing multimodal models even under limited training resources.

A.3.2. Comparison with Retrieval-Augmented and Knowledge-Injection Methods

Notably, our model also achieves clear improvements over retrieval-augmented reasoning (e.g., CARA) and knowledge-injection strategies (e.g., Multimodal-CoT), suggesting that external evidence retrieval or direct knowledge infusion alone is insufficient to address the nuanced reasoning required in VCR. The results indicate that generating and adaptively integrating contextual cues provides a more principled and effective mechanism for strengthening visual commonsense reasoning.

Approach	VCR		
	Q→A	QA→R	Q→AR
SGEITL [10]	76.0	78.0	59.6
PEVL [12]	76.0	76.7	58.6
MSGT [15]	72.2	73.6	53.3
BLIP-2 [3]	75.8	74.3	56.8
ATGAN [11]	72.3	72.9	53.0
CARA [4]	79.3	-	-
Multimodal-CoT [14]	78.8	80.2	63.1
EventLens [5]	82.7	82.7	68.5
mPLUG-Owl3 [13]	82.9	82.1	67.3
Qwen2.5-VL [1]	83.1	81.3	69.4
GPT4RoI [10]	87.4	89.6	78.6
ViP-LLaVA [12]	**87.66**	**89.80**	**78.93**
Ours	85.9	86.1	74.8

Table 2. Performance comparison on VCR dataset.

Approach	VisualCOMET		
	BLEU-2	CIDEr	METEOR
DIVE [6]	13.33	20.26	11.48
Multimodal-CoT [14]	12.13	25.28	12.33
CARA [4]	13.54	46.42	15.58
mPLUG-Owl3 [13]	15.39	40.94	17.11
Qwen2.5-VL [1]	15.91	42.35	16.12
Ours	**16.31**	**48.12**	**17.26**

Table 3. Evaluation results on the VisualCOMET dataset.

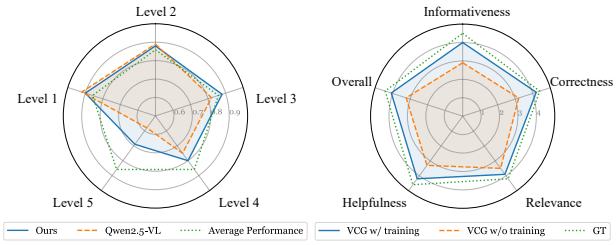


Figure 3. Performance across reasoning difficulty levels and human evaluation of VCG learning effectiveness.

A.3.3. Analysis of Failure Cases

We analyzed the extremes of visual sufficiency scenarios across 300 manually annotated cases. We found that among the 147 cases classified as Levels 1–2 (visually sufficient), there were a total of 14 inference errors. The primary cause of failure was insufficient suppression of external context; the external context generated by the model introduced some noise in visually sufficient scenarios. In the severely visually deficient cases (Levels 4–5), totaling 57 cases, there were 18 inference errors. The Visual Context Generator (VCG) was unable to generate accurate context, which further led to the failure of the gating mechanism.

A.3.4. Inference Cost

The table below details latency profiling on an A100. Although our two-stage pipeline increases total latency (429ms vs. ∼108ms), the magnitude is acceptable. We carefully optimized the trade-off in our experiments to strictly control latency overhead while maximizing reasoning accuracy.

We use a resampler to compress context into 32 learned tokens to minimize latency. This fixed budget confirms gains stem from context quality and gating, not text length.

Method	Backbone Model	Prompting Strategy	External Knowledge	Inference Latency
mPLUG-Owl3	Qwen2 (7B) Memory: 18.4 GB	Identical QA Prompt	None (Implicit)	Single-pass 108ms
Qwen2.5-VL	Qwen2.5 (7B) Memory: 18.1 GB	Identical QA Prompt	None (Implicit)	Single-pass 129ms
Ours	**Qwen2 (7B) Memory: 18.9GB**	**Identical QA Prompt**	**Generated context**	**Two-stage 286ms + 143ms**

Table 4. Performance across reasoning difficulty levels and human evaluation of VCG learning effectiveness.

A.3.5. Impact of Visual Context Generation Learning

The human evaluation results in Fig. 3 (right) demonstrate the effectiveness of the proposed VCG learning strategy in enhancing visual context generation and reducing hallucinations introduced during the generation process. Specifically, We randomly sampled 300 instances from the VisualCOMET dataset, and five human evaluators rated the generated contextual cues on five dimensions-Correctness, Informativeness, Relevance, Helpfulness, and Overall-using a 1-5 scale. Three variants were evaluated: Ground Truth (GT), VCG w/ learning, and VCG w/o learning. VCG w/ learning consistently outperforms VCG w/o learning across all metrics, achieving scores notably closer to those of the ground truth. The most substantial improvements are observed in Correctness and Informativeness, indicating that the learned VCG module produces more accurate and semantically enriched contextual descriptions, thereby mitigating to some extent the hallucinations introduced during the visual context generation process.

A.4. Additional qualitative examples

Fig. 4 illustrates how CoVCR generates complementary internal and external contextual cues to bridge semantic gaps between sparse or ambiguous visual inputs and the corresponding ground-truth answers. In the first case, the internal context not only describes the scene-relevant information but, together with external cues such as "*key data*, *a report*,

(A) **Why is [person1] covering her face?**

Ground truth

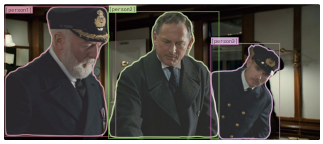
Answer: She seems to be covering her nose from something.

Rationale: [person4]doesn't look as dire as [person1] so it is probably just asmell that only she is smelling.

Reasoning paths

- ***Past Event (What likely happened before?)***
  [Person1] experienced something unsettling nearby, making her feel anxious before the moment shown.
  Loud noises or conflict may have occurred earlier, causing [Person1] to feel frightened.
  The group encountered harsh weather or smoke, prompting [Person1] to shield her face.
  …
- ***Current Event (What is currently happening?)***
  [Person1] is covering her face to protect herself from cold, dust, or ***strong odors***.
  The environment appears tense, making [Person1] react defensively.
  Other people around her seem alert, contributing to her cautious body language.
  …
- ***Future Event (What is likely to happen next?)***
  [Person1] may uncover her face once she feels safer or the discomfort decreases.
  The group might move away from the source of danger or unpleasant conditions.
  Someone nearby could check on [Person1] to ensure she is okay.
  …

(B) **What is [person2] showing to [person1] ?**



Ground truth

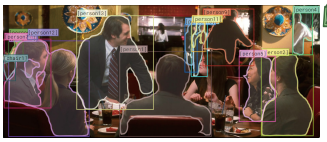
Answer: An important document.

Rationale: [person2] is pointing to something, probably a paper.

Generated contextual prompts

- ***Past Event (What likely happened before?)***
  [Person2] collected ***key data*** and entered the room to deliver a ***report.***
  The crew held a prior meeting or received urgent instructions for review.
  [Person1] and [Person3] discussed operations before [Person2] provided new details.
  …
- ***Current Event (What is currently happening?)***
  [Person2] is presenting a ***document or object*** on the table to officers.
  [Person1] and [Person3], in uniform, are discussing ***a file's contents*** seriously.
  The three individuals are participating in a serious meeting or briefing.
  …
- ***Future Event (What is likely to happen next?)***
  The team may take action, such as reporting or changing the current course.
  [Person1] will likely give instructions based on the ***reviewed document***.
  A detailed planning session might follow with more staff and next steps.
  …

(C) **Why is [person13] smiling?**



Ground truth

Answer: He was recently promoted.

Rationale: He is celebrating with coworkers about his new career promotion.

Reasoning paths

- ***Past Event (What likely happened before?)***
  [Person13] learned about his ***promotion*** earlier today, improving his mood.
  The group planned to meet at the restaurant to ***celebrate recent positive news***.
  [Person13] completed an important work earlier, earning recognition from colleagues.
  …
- ***Current Event (What is currently happening?)***
  Someone is telling an entertaining story, causing the table to react with laughter.
  The group is eating and drinking while maintaining a lively and cheerful atmosphere.
  The restaurant environment feels ***celebratory***, enhancing [Person13]'s happiness during dinner.
  …
- ***Future Event (What is likely to happen next?)***
  They may order dessert or raise a toast to continue ***celebrating the recent promotion***.
  [Person13] might share additional thoughts about future work plans after his ***promotion***.
  Someone could suggest taking a group photo to commemorate this cheerful gathering.
  …

(D) **Why is [person5] looking down?**

Ground truth

Answer: [person5] is looking at the menu.

Rationale: [person5] is in a restaurant and you usually look at a menu before ordering.

Generated contextual prompts

- ***Past Event (What likely happened before?)***
  [Person5] glanced at something on the table earlier, preparing to look at it again.
  She previously ***reviewed the menu*** when deciding what to order for ***dinner***.
  A message or notification may have appeared on her phone earlier.
  …
- ***Current Event (What is currently happening?)***
  [Person5] is looking at an item directly on the table in front of her.
  She appears ***focused on the menu***, checking details or reading a specific section.
  She may be reading a message displayed on her phone near her drink.
  …
- ***Future Event (What is likely to happen next?)***
  [Person5] will likely look up again after finishing ***reading the menu***.
  She may pick up her phone to continue checking a message.
  She may make a dining decision after ***reviewing the menu*** or information she saw.
  …

(E) **Why does [person1] have a photo of [person2]?**

Ground truth

Answer: He is a suspect.

Rationale: [person1] has her ex's listed on a suspect board.

Generated contextual prompts

- ***Past Event (What likely happened before?)***
  [Person1] collected files and assembled the board while ***investigating a specific case***.
  [Person2] had past ties or actions that triggered [person1]'s ***suspicion*** or interest.
  A major incident likely drove [person1] to ***begin this focused investigation***.
  …
- ***Current Event (What is currently happening?)***
  [Person1] stands before an evidence board, closely examining photos and notes.
  [Person2]'s photo is pinned, suggesting they're part of an ongoing ***investigation***.
  The room resembles a workspace filled with instruments and documents.
  …
- ***Future Event (What is likely to happen next?)***
  [Person1] may question or confront [person2] about their role or involvement.
  More ***evidence*** might be added to uncover deeper links between key subjects.
  [Person1] could act on a new lead, ***advancing the investigation further***.
  …

(F) **What is [person6] doing?**

Ground truth

Answer: [person6] is singing a solo in a play.

Rationale: [person6] is positioned in the foreground facing an audience and theother actors standing in the background behind her.

Generated contextual prompts

- ***Past Event (What likely happened before?)***
  The group was previously dancing together, building excitement and energy among everyone present.
  Music started playing, encouraging [Person6] to prepare for a lively performance.
  The crowd had been cheering, motivating [Person6] to step forward enthusiastically.
  …
- ***Current Event (What is currently happening?)***
  [Person6] is performing an expressive dance move with raised arms.
  She appears fully ***immersed in the music***, extending her body during the performance.
  The surrounding group is watching and smiling, supporting her lively dance movement.
  …
- ***Future Event (What is likely to happen next?)***
  [Person6] may continue dancing energetically as the music builds toward a climax.
  The group might join her movement, creating a synchronized or celebratory routine.
  She could finish her gesture and transition into another expressive dance motion.
  …

Figure 4. Case studies illustrating CoVCR's context-enhanced visual reasoning with generated prompts. Key information relevant to answer reasoning is highlighted in bold for clarity.

reviewed document, etc.", supplements important clues beyond the given visual input. Although some generated contextual cues may contain potentially noisy elements, the model is still able to infer the correct answer. These observations indicate that CoVCR's adaptive fusion of internal and external contexts provides effective guidance, enhancing the robustness of visual commonsense reasoning.

A.5. Limitations and Future Work

Although CoVCR shows strong robustness under insufficient visual narratives, several limitations remain. First, the quality of generated context is bounded by the capacity of the visual context generation model. When visual cues are extremely sparse or ambiguous, the generated internal and external context may become generic or misaligned with the scene, introducing noise into reasoning. Second, although the gating-based adapter helps balance contextual contributions, it can struggle with fine-grained visual grounding, such as small objects or subtle interactions that are difficult to capture through text alone. Third, the two-stage training pipeline—pretraining the context generation model followed by finetuning—incurs additional computational cost compared to single-stage approaches. Finally, the external context is limited to textual descriptions; incorporating richer modalities, such as temporal visual cues or structured knowledge, remains an open problem.

Future work can address these limitations in several directions. One direction is to improve the fidelity of generated context by leveraging stronger vision-language models or incorporating temporal signals from video. Another is to explore structured or hybrid context representations, such as event graphs or knowledge bases, to provide more grounded reasoning cues. Improving efficiency through parameter-efficient training is also important. In addition, extending CoVCR to broader scenarios—such as open-world VQA or embodied reasoning—may further test its generality. Finally, studying failure cases and uncertainty in generated context could help determine when contextual augmentation is beneficial.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. 4

[2] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 1

[3] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, pages 19730–19742, 2023. 4

[4] Junzhang Liu, Zhecan Wang, Hammad Ayyubi, Haoxuan You, Chris Thomas, Rui Sun, Shih-Fu Chang, and Kai-Wei Chang. Detecting multimodal situations with insufficient context and abstaining from baseless predictions. In *Proceedings of the ACM International Conference on Multimedia*, pages 8402–8411, 2024. 4

[5] Mingjie Ma, Zhihuan Yu, Yichao Ma, and Guohui Li. Eventlens: Leveraging event-aware pretraining and cross-modal linking enhances visual commonsense reasoning. *arXiv preprint arXiv:2404.13847*, 2025. 4

[6] Jun-Hyung Park, Hyuntae Park, Youjin Kang, Eojin Jeon, and SangKeun Lee. Dive: Towards descriptive and diverse visual commonsense generation. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 9677–9695, 2023. 4

[7] Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. Visualcomet: Reasoning about the dynamic context of a still image. In *Proceedings of the European Conference on Computer Vision*, pages 508–524, 2020. 1

[8] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763, 2021. 1

[9] Qwen Team. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024. 1

[10] Zhecan Wang, Haoxuan You, Liunian Harold Li, Alireza Zareian, Suji Park, Yiqing Liang, Kai-Wei Chang, and Shih-Fu Chang. Sgeitl: Scene graph enhanced image-text learning for visual commonsense reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, pages 5914–5922, 2022. 4

[11] Shaojuan Wu, Kexin Liu, Jitong Li, Peng Chen, Xiaowang Zhang, and Zhiyong Feng. Temporal-based graph reasoning for visual commonsense reasoning. *Knowledge-Based Systems*, page 113214, 2025. 4

[12] Yuan Yao, Qianyu Chen, Ao Zhang, Wei Ji, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Pevl: Position-enhanced pre-training and prompt tuning for vision-language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 11104–11117, 2022. 4

[13] Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mPLUG-owl3: Towards long image-sequence understanding in multi-modal large language models. In *Proceedings of the International Conference on Learning Representations*, 2025. 4

[14] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *Transactions on Machine Learning Research*, 2024. 4

[15] Jian Zhu, Hanli Wang, and Bin He. Multi-modal structure-embedding graph transformer for visual commonsense reasoning. *IEEE Transactions on Multimedia*, pages 1295–1305, 2023. 4