Knowledge Bridger: Towards Training-Free Missing Modality Completion – Supplementary Materials –

A. Prior Knowledge Rules and Prompts

We report the knowledge extraction rules used in our method and the prompts used in the knowledge-driven generation respectively.

A.1. Prior Knowledge

To simplify the research problem, we utilize only a set of basic instructions to direct the large language model (LMM) focus toward the desired multimodal content extraction. For the general domain, we predefine the extracted knowledge to include major objects, the quantity of each object, and their corresponding attributes and styles. Given the strong reasoning capabilities of LMM in the general domain, it is not necessary to differentiate between input modalities. However, to mitigate hallucinations generated by the LMM, we limit the number of major objects. In our experiments, extracting between 5 to 7 objects show optimal.

General Domain Rules

Understand the given *[input-format]* to extract the following information:
- Identify the top *[object-numbers]* objects by their specific names (e.g., 'man', 'woman' instead of 'person').
- Specify the count of each identified object.
- Describe attributes for each object in detail.
- Summarize the style of the *[input-format]*.

In the medical domain, it is necessary to pay attention to the distinctions between different modalities. This is because LMM may not inherently comprehend the knowledge required for out-of-domain (OOD) scenarios, necessitating clear identification of the modality being processed and the content to be understood. In our approach, for X-rays, we instruct the LMM to focus on anatomical structures, clinical significance, abnormal findings, and report generation.

Medical Domain Rules (X-ray)

This is a chest X-ray image. Please follow these steps for a comprehensive analysis:
- Describe the main anatomical structures visible in the image, such as the lungs, heart, and trachea.
- Identify any abnormalities present, such as opacities, nodules, or effusions, and describe their characteristics.
- Explain the potential clinical significance of any abnormalities noted.
- Summarize the findings and draft a detailed clinical report based on your observations.

For reports, in addition to the information extracted from X-rays, we direct the LMM to consider the locations of the anatomical structures mentioned and any additional characteristic details present in the report.

Medical Domain Rules (Report)

Given the following clinical report, analyze and identify specific visual details that would correspond to the described findings on a chest X-ray. Follow these steps:
- Identify the main anatomical structures mentioned in the report and locate them on a chest X-ray.
- Highlight the abnormalities or specific findings described in the report.
- Describe the characteristics (e.g., size, shape, density) of these abnormalities.
- Relate these characteristics to potential clinical conditions.
- Summarize your analysis with a list of visual features expected in the X-ray.

A.2. Knowledge Extraction with Chain-of-Thought

We employ an LMM with Chain-of-Thought (CoT) [16] reasoning to extract knowledge according to the aforementioned rules. This approach helps reduce the computational strain associated with long-problem reasoning and enhances the overall accuracy of problem-solving. For the general

domain, our final instruction prompt is designed as follows:

General Domain Knowledge Extraction

Role: SYSTEM
Content: You are a helpful assistant in understanding images and texts and you can extract very important and accurate information from them.

Role: USER
Content:
Instruction
Your task is to understand the user's inputs and extract the related information following the instruction:
{ RULES }

{ User Input }
Please process each point step by step.

Role: ASSISTANT
Content: ...

For medical domain, we have:

Medical Domain Knowledge Extraction

Role: SYSTEM
Content: You are a very experienced radiologist.

Role: USER
Content:
Instruction
The following are some chest x-ray image and report examples. Your task is to understand the images and reports, and extract the important information based on the following questions.
Examples
Example 1:
Chest X-ray Image: *[Image 1]*.
Clinical report: *[Report 1]*.

Example 2:
Chest X-ray Image: *[Image 2]*.
Clinical report: *[Report 2]*.

Query
{ RULES }

{ User Input }
Please process each point step by step.

Role: ASSISTANT
Content: ...

Next, we require the LMM to integrate the CoT results according to the following instructions:

Integrating the CoT Results

Role: USER
Content:
Instruction
- You have to integrate the previous result into a structure format.
- Use precise nouns and avoid general terms; each object should be accurately named.

Return Format
The output must be in JSON format as follows:
{ return-format }

Role: ASSISTANT
Content: ...

We strictly require the LMM to return structured information in the following format:

General Domain Return Format

{
"objects": ["Obj. 1", "Obj. 2", ...],
"numbers": {
"Obj. 1": 2,
"Obj. 2": 1,
...
},
"attributes": {
"Obj. 1": "Description of attributes here.",
...
},
"style": "Description of style here."
}

Medical Domain Return Format

Structured Analysis
1. **Anatomical Structures**:
- Lungs: [Left Upper Lobe: Normal/Abnormal], [Right Lower Lobe: Normal/Abnormal]
- Heart: [Normal/Abnormal]
- Trachea: [Normal/Abnormal]

2. **Type of Abnormality**:
- Identified Abnormality: [e.g., opacity, nodule, effusion]
- Characteristics: [e.g., size: 2 cm, shape: round, border: well-defined/ill-defined, density: high]

3. **Distribution and Location**:
- Side: [Unilateral/Bilateral]
- Location: [Upper/Lower/Middle lobe]
- Extent: [Localized/Diffuse]

4. **Clinical Implication**:
- Possible Diagnosis: ['No Finding', 'Enlarged Cardiomediastinum', 'Cardiomegaly', 'Lung Opacity', 'Lung Lesion', 'Edema', 'Consolidation', 'Pneumonia', 'Atelectasis', 'Pneumothorax', 'Pleural Effusion', 'Pleural Other', 'Fracture', 'Support Devices']
- Recommended Action: [Further imaging, clinical follow-up, etc.]

After extracting the aforementioned structured information, we employ LMM to transform these knowledge into the form of a knowledge graph. To simplify the process, we represent relationships on the graph using a triplet structure (nodes and edges):

Building Knowledge Graphs

Instruction
Your task is to analyze the provided *[input-type]* and extract **exactly *[numbers-of-relationships]* distinct relationships** to build a knowledge graph. Each relationship should be structured as (Head, Relation, Tail), focusing on **clear, direct relations** (e.g., "causes," "is a part of," "describes," etc.).
{ User Input }
Return Format
The output must be in JSON format as follows:

[
{
"head": ...,
"relation": ...,
"tail": ...
},
...
]
Please process each point step by step.

A.3. Knowledge-driven Generation

We employ the knowledge graphs extracted by the aforementioned process as input and employ the LMM to process this information to generate meaningful descriptions of missing modalities. Various modality generators are then utilized to produce the missing information based on these descriptions. For missing image, predictions are based on observable text:

General Domain Image Generation

Instruction
- Expand the basic sentence to *[num-prompts]* high-quality description based on previous analysis and structured data.
- Each new prompt should emphasize different object attributes or scene details.
- **Basic Sentence**: *[text-content]*
[Knowledge Graphs]
Output Format
Output the prompt format must in JSON:

[
"description 1",
.....,
"description K"
]

Medical Domain Image Generation

Instruction
Using the following structured analysis, this information is organized to generate *[num-prompts]* meaningful clinical description:

[Knowledge Graphs]

{ User Input }

Output Format
Output the prompt format must in JSON:

[
"description 1",
....,
"description K"
]

For missing text, the same method is applied to generate descriptions of the missing content, which are then refined by the LMM to produce the required missing text.

A.4. Knowledge-based Ranking Pseudo-code

The pseudo-code for the ranking process is shown in Alg. 1.

B. Implementation Details

GPUs Details. We conduct all experiments on the PyTorch 2.4.0 [11] platform, running on Ubuntu 20.04 LTS utilizing 4 GPUs (NVIDIA GeForce RTX 4090 with 24 GB of memory).

Deploy Efficient Large Multimodal Model. We deploy the Qwen-VL[13] large model using vLLM [7]. vLLM is an LLM serving system that achieves (1) near-zero waste in KV cache memory and (2) flexible sharing of KV cache within and across requests to further minimize memory usage. We deploy versions with 2B[1], 7B[2], and 72B parameters. Specifically, due to hardware constraints, we utilize the 72B quantized version with Int-8 precision available on Hugging Face: Qwen2-VL-72B-Instruct-GPTQ-Int8[3]. For all versions, we maintain an 8K context window length and support a maximum of four image queries. For each query, we set the maximum number of tokens to 512 and use a temperature of 0.1.

Generators Settings. For image reconstruction, we apply Stable Diffusion XL (SDXL) 1.0 [12] as the restoration module for general domains. SDXL 1.0 is an advanced text-to-image diffusion model that can generate images according to a given prompt. Additionally, for the restoration of chest X-ray modality, we use Cheff [15], a cascaded chest X-ray latent diffusion model. By default, we generate 5 candidates for the missing modality during the generation process.

Missing Modality Simulation In our setting, we conduct the missing rate $\eta = \{0.3, 0.5, 0.7\}$ to simulate the missing modality scenario during training. Specifically, we calculate the number of missing samples in different datasets under a given missing rate and then randomly mark the text and image modalities of these samples as missing with a probability of 0.5. To ensure the reproducibility of the experiment, we perform multiple simulations using the same set of missing samples. Finally, we retrain the baseline model, which was initially trained on complete modalities, using the data with imputed missing modality and report the performance across various metrics.

Algorithm 1: Ranking Module. python-style pseudocode

```
# f_a(·), f_c(·), f_b(·): the adjacency
 matrix, CLIP's embedding, and
 BLIP's embedding of the given
 modality, respectively.
# cos_graph(·,·), cos(·,·): Graph
 similarity and embedding
 similarity.
# C: Missing generation
 candidates.
# A: Available modality.

# Quality Scores
QS = []
for c in C:  # load a candidate.
   # Computing the graph similarity
   graph_simi = cos_graph(f_a(A), f_a(c))
   # Computing the embedding
    similarity by CLIP
   clip_simi = cos(f_c(A), f_c(c))
   # Computing the embedding
    similarity by BLIP
   blip_simi = cos(f_b(A), f_b(c))
   score = graph_simi + clip_simi +
    blip_simi
   QS.append(score)
# Ranking.
max_c = QS.index(max(QS))
return C[max_c]
```

[1] https://huggingface.co/Qwen/Qwen2-VL-2B-Instruct
[2] https://huggingface.co/Qwen/Qwen2-VL-7B-Instruct
[3] https://huggingface.co/Qwen/Qwen2-VL-72B-Instruct-GPTQ-Int8

C. More Experimental Results

C.1. Table of Quantitative Results

We present additional quantitative analysis results, as shown in Table 1.

Missing Rate η	0.5		
Method	F1	AP	SS
Baseline (complete)	F1: 57.0	AP: 75.7	SS: -
Baseline (remove missing)	42.1	63.9	-
MPMM [8] (CVPR'23)	42.8	64.2	-
MPLMM [5] (ACL'24)	<u>43.3</u>	<u>65.9</u>	-
MMIN [17] (ACL'21)	34.8	60.1	17.0
DiCMoR [14] (CVPR'23)	37.6	63.8	<u>17.7</u>
Ours (Qwen-VL-2B)	46.6	67.1	21.1
Ours (Qwen-VL-7B)	**50.9**	**70.2**	**21.5**
Δ Complete Baseline	-6.1	-5.5	-
Δ SOTA	+7.6	+4.3	+3.8

Table 1. **Quantitative results (%) on IU X-ray datasets. Bold** denotes the best results and <u>underline</u> denotes the second-best. SS (%) refers to the average similarity score, which is used to assess the generation quality of imputation-based methods. A higher score indicates better quality. '-' indicates that the metric is not applicable. All results are reproduced using the officially released code.

C.2. Table of Ablation Study

We present complete results of the ablation study in Table 2.

C.3. More Modalities Results

The proposed method primarily focuses on image and text modalities to facilitate the evaluation of the proposed method, as these modalities are well-supported by the community and computationally efficient when using large vision-language models. However, our approach imposes no constraints on the modality encoders, allowing it to be easily generalized to other modalities. To validate this, we conducted experiment on a multimodal sarcasm detection dataset [2], which includes audio, vision, and text modalities. Using Unified-IO [1] as the backbone and keeping other configurations unchanged, our method demonstrated effectiveness even when extended to these modalities, as shown in Table 3.

D. Visualization Analysis

D.1. Completion Results

We present additional completion results, as shown in Fig. 1, 2, and 3. In the general domain, our knowledge modeling module emphasizes understanding the quantity of objects, their attributes, and the contextual environment. The results in Fig. 1 indicate that our method more closely resembles the original missing modality compared to direct generation approaches. In the medical domain, incorporating knowledge of different lesions allows the LMM to comprehend the relationships between various regions in chest X-rays and the content described by the modality. Figs. 2 and 3 demonstrate that our method offers a more reliable strategy for missing data completion than direct generation.

D.2. Intermediate Results

We present some intermediate results as shown in Figs. 5 and 6.

D.3. Knowledge-based Ranking Results

We present partial results of knowledge-based ranking as shown in Fig. 4. Here, "available [modality]" indicates that the modality is visible, and "QS" represents the quality score of the completed missing modality combined with the observed modality. The results demonstrate that our proposed knowledge-based ranking module can effectively select relatively reasonable generated outcomes.

E. Hallucinations

The hallucinations of LMM [3] refer to instances when an AI model generates content that is factually incorrect, misleading, or unsubstantiated. In our approach, hallucinations primarily stem from the limitations of the training data. We advocate for a training-free method to achieve MMC, which has the advantage of being easily deployable across various domains with minimal input of relevant domain knowledge. However, the drawback is that the lack of task-specific training can result in the model having less "common sense" compared to models trained on extensive datasets. As illustrated in Fig. 7, the absence of common sense in our approach may lead to results that deviate from expected cognitive outcomes. In recent years, RAG (Retrieval-Augmented Generation) [6, 9] has been regarded as an effective technique for mitigating hallucinations in large models. This technique provides the model with external truths, thereby reducing hallucinations during the reasoning process. In future work, we plan to incorporate RAG to enhance the robustness of our approach.

F. Limitations

F.1. More Modalities

Our method focuses exclusively on image and text modalities, leaving its performance on other modalities, such as speech and depth, yet to be explored. The approach emphasizes the automatic extraction of inter-modal knowl-

Missing Rate η	MM-IMDb									IU X-ray								
	0.3			0.5			0.7			0.3			0.5			0.7		
Variants	F1	AP	SS	F1	AP	SS	F1	AP	SS	F1	AP	SS	F1	AP	SS	F1	AP	SS
Baseline (Qwen-VL-7B)	54.7	60.9	33.5	54.9	61.3	32.7	55.2	61.8	32.3	53.6	73.9	22.6	50.9	70.2	21.5	46.3	70.5	19.8
w/o Knowledge Modeling	-1.2	-3.3	-7.0	-1.5	-4.1	-8.8	-1.3	-3.6	-8.8	-12.1	-21.6	-10.7	-13.3	26.8	-11.6	-17.5	-29.2	-13.7
+ Random Ranking	-1.7	-3.5	-7.2	-1.6	-4.3	-9.0	-1.6	-4.1	-9.9	-13.9	-26.8	-11.4	-14.7	-28.1	-12.4	-19.3	-31.8	-15.0
Random Ranking	-0.5	-2.6	-0.6	-0.4	-2.8	-0.7	-0.5	-2.7	-0.6	-2.9	-6.3	-3.8	-3.1	-6.9	-4.3	-3.8	-7.1	-4.7
w/o Knowledge Ranking	-0.1	-0.4	-0.2	-0.2	-1.9	-0.4	-0.2	-0.8	-0.1	-1.3	-3.3	-1.5	-2.1	-5.4	-3.6	-1.9	-5.7	-2.1
w/o Semantic Ranking	-0.3	-1.4	-0.2	-1.4	-0.2	-0.1	-0.2	-1.0	-0.3	-0.9	-0.4	-0.7	-1.3	-2.3	-1.1	-2.4	-3.3	-1.6

Table 2. **The impact of various components.** We report the comparison results between different combinations and the baseline.

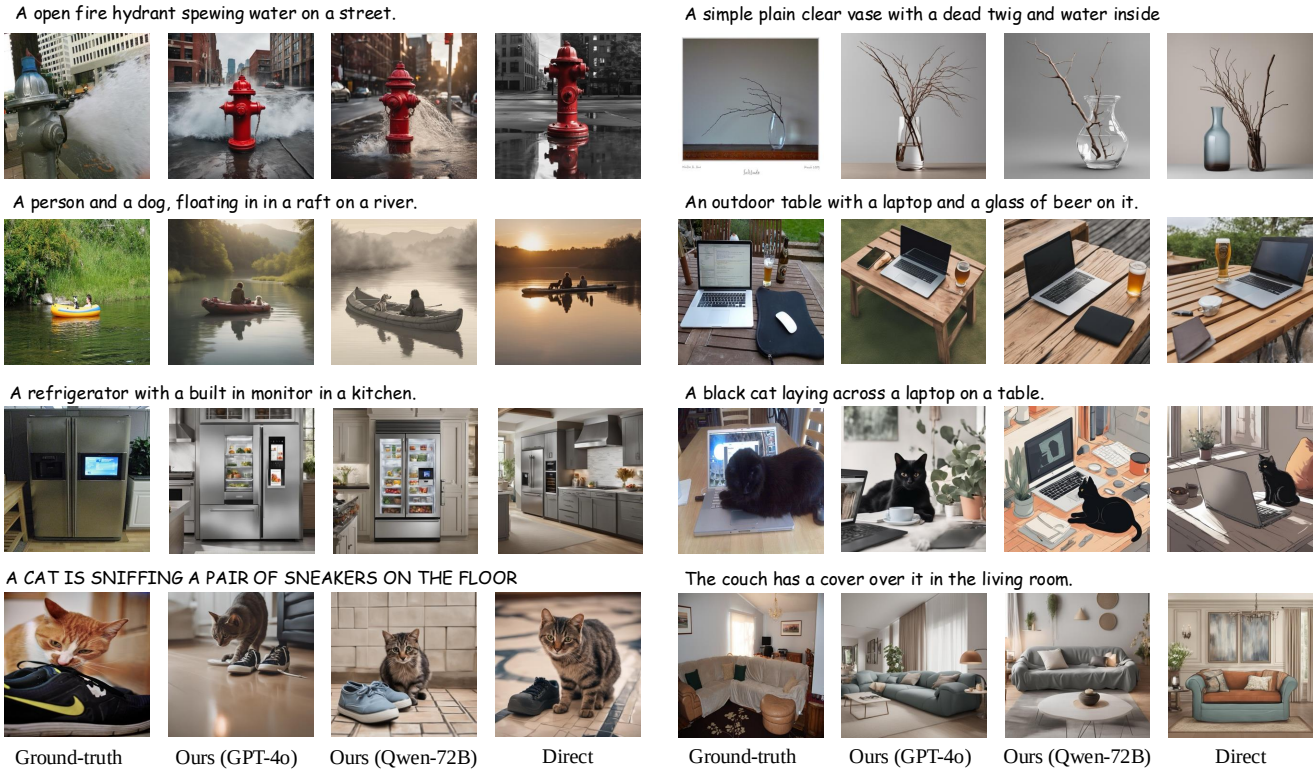

Figure 1. **Visualization analysis.** We present the results of image completion on the COCO dataset. The first and fifth columns display the ground truth images, while the fourth and eighth columns show images generated directly by LMM using the available textual modality. The remaining columns illustrate the outcomes produced by our method, which employs LMM of varying scales.

Missing Rate η	0.3			0.5		
Method	F1	AP	SS	F1	AP	SS
Baseline (complete)	F1: 62.4 | mAP: 64.7| SS: -					
Baseline (remove missing)	57.1	59.3	-	53.8	54.6	-
DiCMoR (CVPR'23)	55.5	57.6	11.7	51.3	52.9	10.2
Ours (Unified-IO 7B)	58.3	60.1	13.3	54.7	55.8	11.4

Table 3. Quantitative results (%) on sarcasm datasets.

edge and the completion of missing modalities through domain knowledge. However, this focus on a limited set of modalities limits its generalizability and adaptability in real-world applications where multi-modal data often involves various types of sensory inputs. Thus, in the future, adaptation to other modalities is possible by defining a more comprehensive modality knowledge and expanding the learning framework to accommodate these new modalities. Some promising works [4, 10] show that one modality, such as image or text, can be connected to any other modality, paving the way for more inclusive and versatile multimodal systems that handle diverse data types with high ef-

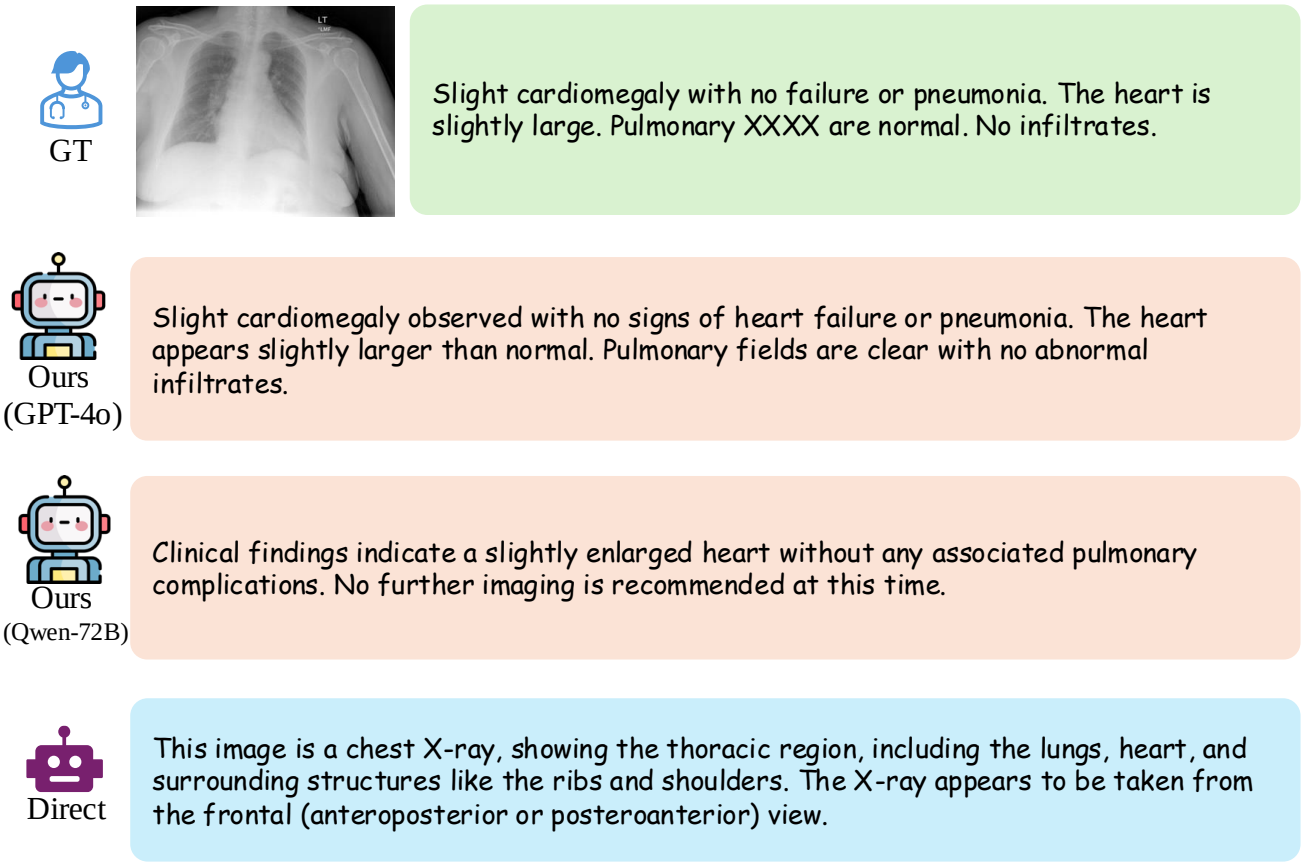

Figure 2. **Visualization analysis.** We present the results of completing missing reports based on the X-ray modality from the IU X-ray dataset.

ficacy.

F.2. More Tasks and Metrics

Additionally, we observe that while our method enhances classification performance under a high missing rate (e.g., 0.7), it paradoxically results in a decrease in the similarity scores of the completed modalities. This suggests that although the model performs well in reconstructing missing data for classification tasks, the semantic alignment and quality of the generated modalities may still require significant refinement. Addressing these limitations presents an opportunity to improve the balance between classification accuracy and modality similarity. Therefore, there remains substantial potential for further exploration to develop more robust generation and ranking strategies in the future. These improvements could include incorporating advanced similarity-preserving techniques and exploring diverse evaluation metrics to assess the completeness and coherence of generated data across different tasks.

References

[1] Jiasen Lu et. al. Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action. In *CVPR*, pages 26439–26455, 2024. 5

[2] Santiago Castro et. al. Towards multimodal sarcasm detection. In *ACL*, pages 4619–4629, 2019. 5

[3] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024. 5

[4] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *CVPR*, pages 15180–15190, 2023. 6

[5] Zirun Guo, Tao Jin, and Zhou Zhao. Multimodal prompt learning with missing modalities for sentiment analysis and emotion recognition. In *ACL*, pages 1726–1736, 2024. 5

[6] Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. Lightrag: Simple and fast retrieval-augmented generation. *arXiv preprint arXiv:2410.05779*, 2024. 5

[7] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for

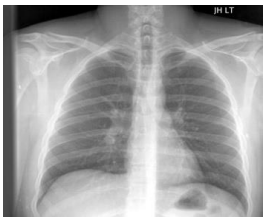

Figure 3. **Visualization analysis.** We present the results of completing missing reports based on the X-ray modality from the IU X-ray dataset.

large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626, 2023. 4

[8] Yi-Lun Lee, Yi-Hsuan Tsai, Wei-Chen Chiu, and Chen-Yu Lee. Multimodal prompting with missing modalities for visual recognition. In *CVPR*, pages 14943–14952, 2023. 5

[9] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020. 5

[10] Yuanhuiyi Lyu, Xu Zheng, Jiazhou Zhou, and Lin Wang. Unibind: Llm-augmented unified and balanced representation space to bind them all. In *CVPR*, pages 26752–26762, 2024. 6

[11] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *NIPS*, 32, 2019. 4

[12] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *ICLR*, 2024. 4

[13] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 4

[14] Yuanzhi Wang, Zhen Cui, and Yong Li. Distribution-consistent modal recovering for incomplete multimodal learning. In *ICCV*, pages 22025–22034, 2023. 5

[15] Tobias Weber, Michael Ingrisch, Bernd Bischl, and David Rügamer. Cascaded latent diffusion models for high-resolution chest x-ray synthesis. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 180–191. Springer, 2023. 4

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

[17] Jinming Zhao, Ruichen Li, and Qin Jin. Missing modality imagination network for emotion recognition with uncertain missing modalities. In *ACL*, pages 2608–2618, 2021. 5

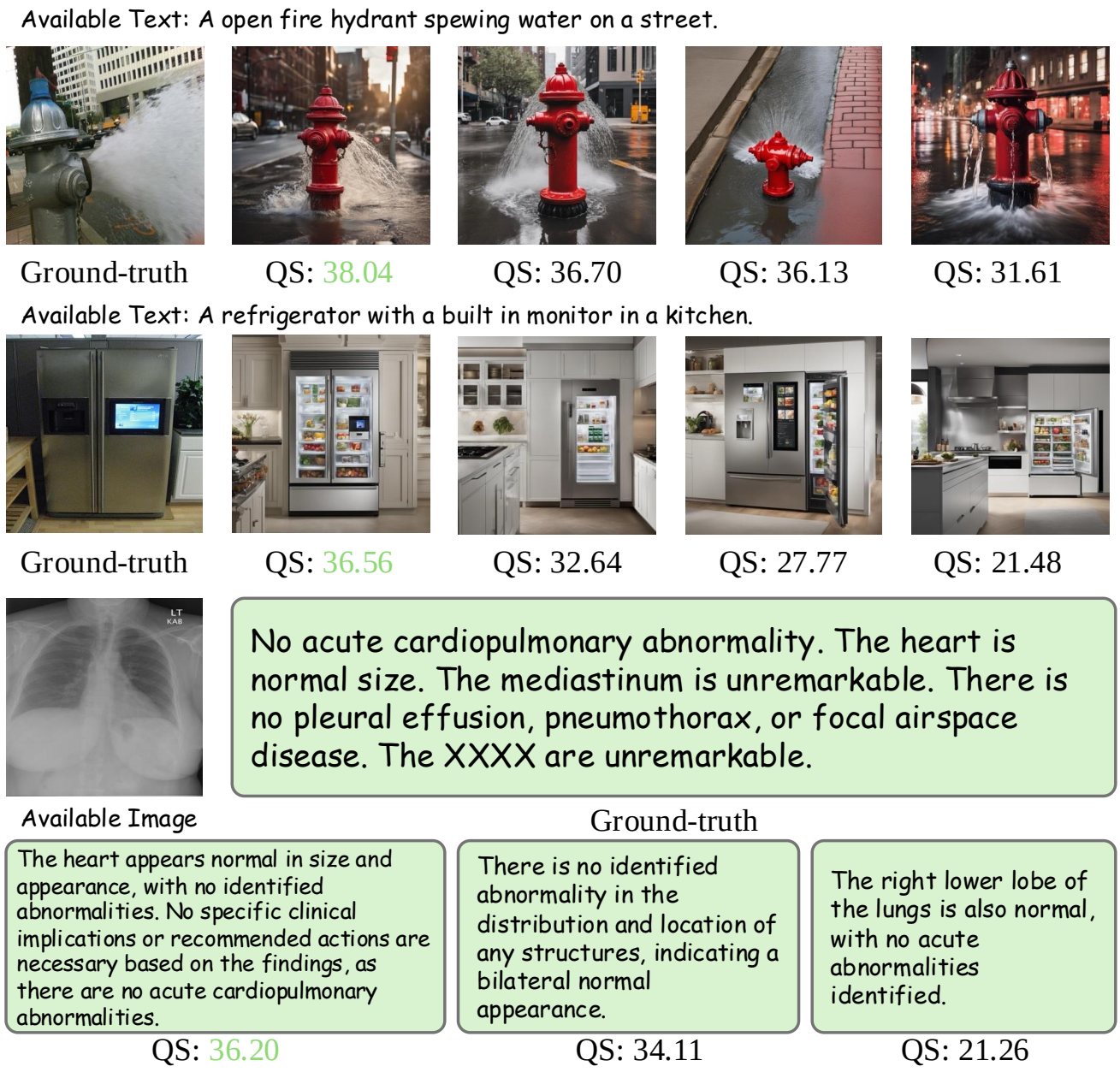

Figure 4. **Visualization of Knowledge-based Ranking.** We present the results of knowledge-based ranking.

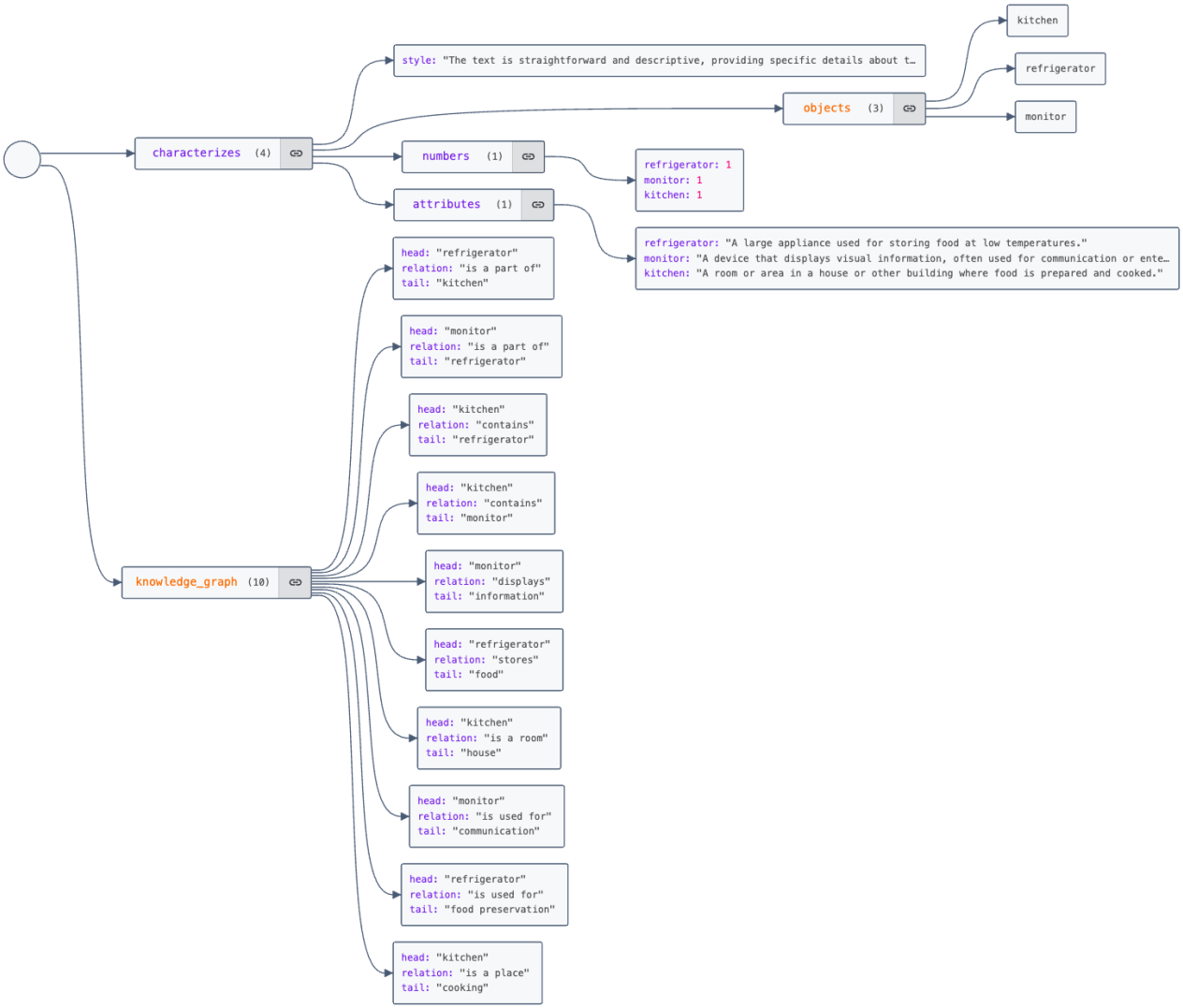

Figure 5. **Intermediate results.** We present the intermediate results extracted by our method, referred to as knowledge.

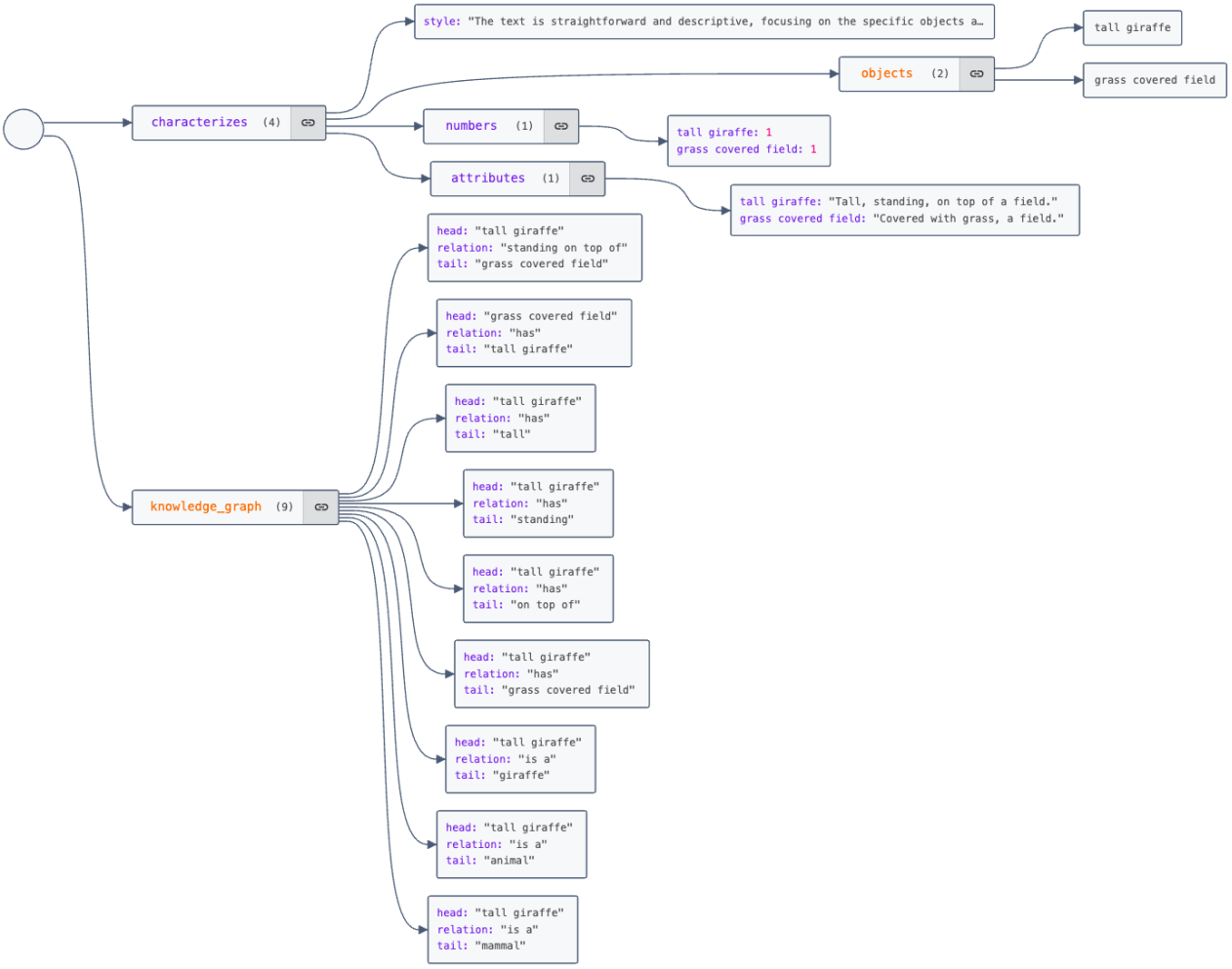

Figure 6. **Intermediate results.** We present the intermediate results extracted by our method, referred to as knowledge.

Figure 7. Hallucinations analysis.