Beyond Embeddings: The Promise of Visual Table in Visual Reasoning

Yiwu Zhong[*1], **Zi-Yuan Hu**[*12], **Michael R. Lyu**[1], **Liwei Wang**[†1]
[1]The Chinese University of Hong Kong [2]Shanghai AI Laboratory
yiwuzhong@cuhk.edu.hk {zyhu22,lyu,lwwang}@cse.cuhk.edu.hk

Abstract

Visual representation learning has been a cornerstone in computer vision, involving typical forms such as visual embeddings, structural symbols, and text-based representations. Despite the success of CLIP-type visual embeddings, they often lack access to world knowledge critical for visual reasoning. In this work, we propose **Visual Table**, a novel form of visual representation tailored for visual reasoning. Visual tables are constructed as hierarchical descriptions of visual scenes, featuring a scene description and multiple object-centric descriptions covering categories, attributes, and knowledge. Thanks to the structural and textual formats, visual tables offer **unique properties** over mere visual embeddings, such as explainability and controllable editing. Furthermore, they deliver instance-level world knowledge and detailed attributes that are essential for visual reasoning. To create visual tables, we develop a generator trained on the dataset with collected, small-scale annotations. Extensive results on **11** visual reasoning benchmarks demonstrate that the generated visual tables **significantly** outperform previous structural and text-based representations. Moreover, they **consistently** enhance state-of-the-art multi-modal large language models across diverse benchmarks, showcasing their potential for advancing visual reasoning tasks. Our code is available at https://github.com/LaVi-Lab/Visual-Table.

1 Introduction

Visual representation learning has been a fundamental and long-standing topic in computer vision. At the early stage, the learning was supervised by expensive human-annotated labels (Krizhevsky et al., 2012; Szegedy et al., 2015; He et al., 2016). This paradigm recently evolved to learn visual embeddings by aligning image-text pairs from the Internet (Radford et al., 2021; Jia et al., 2021). Beyond visual embeddings, symbolic and structured visual representations (*e.g.*, scene graph) (Xu et al., 2017; Krishna et al., 2017) exhibited advantages across domains, such as vision-language tasks (Teney et al., 2017; Hudson and Manning, 2019a; Zhong et al., 2020), video and 3D scene understanding (Yang et al., 2023a; Armeni et al., 2019; Wald et al., 2020), and robotics (Rana et al., 2023; Gu et al., 2023; Kalithasan et al., 2023). More recently, some works have strived to convert visual scenes into text-based representations (*e.g.*, image captions) (Hu et al., 2022; Yang et al., 2022; Shao et al., 2023; Khademi et al., 2023), triggering the reasoning capability of large language models (LLMs) (Ouyang et al., 2022; Zhang et al., 2022; Touvron et al., 2023; Chiang et al., 2023).

Among these visual representations, CLIP-type visual embeddings (Radford et al., 2021), learned from image-text pairs, have dominated many vision tasks. Their success can be attributed to robust generalization in encoding visual attributes (*e.g.*, visual appearance (Yang et al., 2023b; Pratt et al., 2023; Yan et al., 2023), visual relations (Zhao et al., 2023; Li et al., 2024; Zhong et al., 2023; Momeni et al., 2023)). However, lacking external world knowledge (*e.g.*, object affordance, background knowledge of named entities), these visual embeddings lead to sub-optimal performance in complex reasoning tasks. Consider the example at the left top of Fig. 1. While CLIP embeddings can effectively capture visual attributes (*e.g.*, recognizing a person with a beard), they still struggle to answer questions that necessitate critical world knowledge beyond the image itself (Marino et al., 2019; Schwenk et al., 2022; Chen et al., 2023c; Yue et al., 2023).

To this end, we propose **Visual Table**, a novel form of visual representation designed for visual reasoning. As shown in the left-bottom of Fig. 1, a visual table is presented as hierarchical text descriptions for visual scenes, and comprises a scene de-

[*]equal contributions, † corresponding author.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 6876–6911
November 12-16, 2024 ©2024 Association for Computational Linguistics

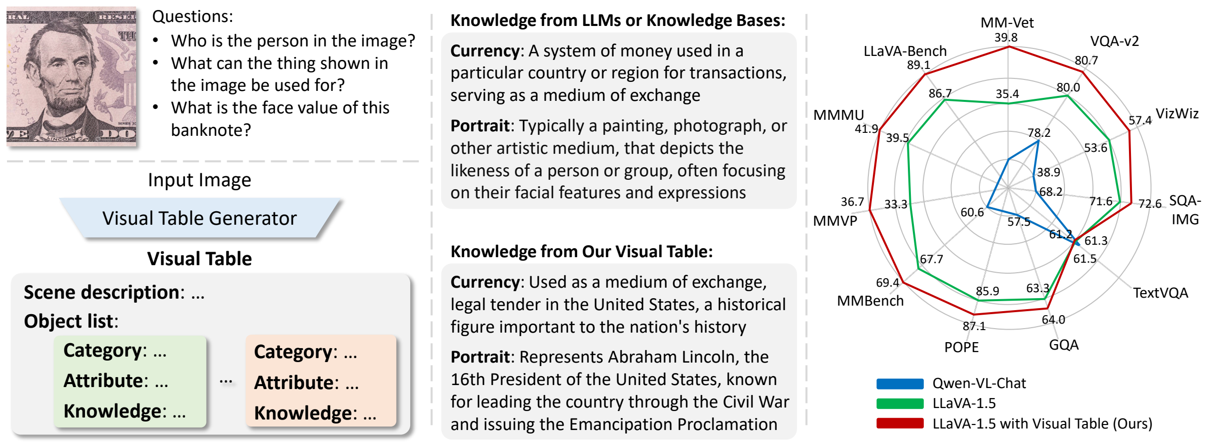


Figure 1: **Left Top:** An example that requires world knowledge to answer the questions about the image. **Left Bottom:** Our proposed visual tables represent thorough visual content in structured text. **Middle:** Visual tables provide instance-level knowledge in considering specific visual instances. **Right:** The generated visual tables can consistently enhance the state-of-the-art multi-modal large language models across diverse benchmarks.

scription and multiple object-centric descriptions, covering object categories, object attributes, and object knowledge. Through structural and textual formats, visual tables offer unique properties beyond mere visual embeddings. First, **textual formats** can be seamlessly processed by LLMs without the need for training, while also facilitating interpretation and interaction with humans. Second, **hierarchical structures** support controllable table editing, enabling operations such as removal and insertion, and allowing customization of visual representation for different downstream tasks. Third, by capturing visual scenes in **rich details**, such as the object-centric descriptions in free-form language covering attribute and knowledge, visual tables hold the potential to enhance visual reasoning across various domains, spanning images, videos, 3D scenes, and beyond. Note that **instance-level** object knowledge is generated in consideration of specific visual instances, unlike previous works that retrieve **category-level** knowledge from frozen LLMs or knowledge bases (Yang et al., 2022; Gui et al., 2021; Lin et al., 2022; Shao et al., 2023). For example, in the middle of Fig. 1, the visual table provides the world knowledge customized for the particular object instance, *a banknote of the United States with Abraham Lincoln who is the 16th President and leads the Civil War*, thereby supporting complex visual reasoning.

In this work, we focus on proof-of-concept for visual tables and validate its design and efficacy in image reasoning tasks. To create our designed visual tables, a straightforward solution is to collect human annotations and use them to train a generator. Despite high quality, it is costly in terms of the training required for human annotators and the label normalization during post-processing. To reduce these burdens, we adopt an alternative solution leveraging foundation models with near-human capabilities (OpenAI, 2023; Team et al., 2023; Anthropic, 2024). Specifically, we design prompts to ensure the annotations closely align with our specifications. After collecting a small scale of annotations, *i.e.*, 61K visual tables, we train a generator capable of automatically producing visual tables for arbitrary images.

To validate the effectiveness of visual tables as visual representations, we adopt **11 diverse benchmarks** that require an understanding of object attributes and knowledge. We seek to answer two key questions: (1) Do visual tables represent visual scenes more effectively than traditional structural and text-based representations (*e.g.*, scene graphs, image captions)? (2) Can visual tables enhance the performance of existing multi-modal large language models (MLLMs), which already demonstrated incredible results on visual reasoning (Dai et al., 2023; Liu et al., 2023a; Bai et al., 2023)? Our extensive experiments reveal that visual tables **significantly** outperform previous structural and text-based representations across various benchmarks. Furthermore, visual tables **consistently** improve the performance of state-of-the-art (SOTA) MLLMs (*e.g.*, LLaVA-1.5 (Liu et al., 2023a)), as shown at the right of Fig. 1. These findings validate that visual tables can facilitate visual reasoning by serving as general, robust visual representations.

Our contributions are summarized as follows:

(1) We propose Visual Table, a new form of visual representation organized in structural text. It offers unique benefits beyond visual embeddings — explainability, controllable editing, and instance-level knowledge. (2) We introduce a new dataset with 61K visual table annotations, and present a generator that can produce high-quality visual tables for any input images. The dataset, together with our visual table generator, can be further exploited by future research. (3) Extensive experiments show that visual tables, working as generalizable representations, largely outperform previous structural, text-based representations and consistently improve the SOTA MLLMs across benchmarks.

2 Related Work

Visual Representation Learning. Early works focus on training image classifiers using labor-intensive image labels (Krizhevsky et al., 2012; Simonyan and Zisserman, 2015; Szegedy et al., 2015; He et al., 2016). As an augmentation, object and attribute labels are used to enhance visual representation (Anderson et al., 2018; Zhang et al., 2021). To reduce the annotation cost, self-supervised learning (He et al., 2020; Chen et al., 2020; Grill et al., 2020; He et al., 2022) is proposed to match the visual representation of different views from the same image. Moving forward, vision-language pretraining (Radford et al., 2021; Jia et al., 2021; Gu et al., 2021; Zhong et al., 2022; Li et al., 2022) is proposed to match web-collected visual-text pairs, exhibiting generalizable capability on diverse visual recognition tasks. Compared to these visual embeddings, our visual tables additionally encode **world knowledge** critical for visual reasoning, support **interpretability** to humans and LLMs, and enable **controllable editing**.

Beyond implicit embeddings, structural and symbolic representations, such as image scene graphs (Xu et al., 2017; Zellers et al., 2018; Tang et al., 2020; Shi et al., 2021; Zhong et al., 2021), have attracted significant attention. These works aim to abstract visual scenes into concise representations, demonstrating special benefits in diverse domains, including vision-language modeling (Yu et al., 2021; Pan et al., 2022; Mitra et al., 2023; Herzig et al., 2023) and various downstream tasks (Hudson and Manning, 2019a; Zhong et al., 2020; Ji et al., 2020; Hughes et al., 2022; Kalithasan et al., 2023). Resembling the concept of structural representations, visual tables are presented in the hierarchical text, yet deliver **richer semantics** through free-form language.

Another line of works explores text-based visual representation (Hu et al., 2022; Wang et al., 2023b; Yang et al., 2022; Gui et al., 2021; Lin et al., 2022; Shao et al., 2023; Fu et al., 2023b; Khademi et al., 2023; Wang et al., 2022; Hakimov and Schlangen, 2023). These methods typically convert visual inputs into text (*e.g.*, image captions, object tags), then retrieve knowledge from knowledge bases (*e.g.*, Wikipedia (Vrandečić and Krötzsch, 2014), ConceptNet (Liu and Singh, 2004)) and/or frozen LLMs (*e.g.*, GPT3), and finally perform text reasoning using frozen LLMs with in-context examples. Unlike these works, our **single** generator model learns to compress a comprehensive visual knowledge base, **without the need** for manually selecting off-the-shelf vision libraries (*e.g.*, image captioners, object detectors), external knowledge sources, or high-quality in-context examples.

Multi-modal Large Language Models. MLLMs harness LLMs to empower reasoning on multi-modal tasks, typically on visual question answering (VQA) (Li et al., 2023b; Fu et al., 2023a; Yin et al., 2023; Alayrac et al., 2022; Dai et al., 2023; Zhu et al., 2023; Chen et al., 2023a; Bai et al., 2023; IDEFICS, 2023; Huang et al., 2024; Li et al., 2023a). These methods usually learn layers that connect visual encoder (Radford et al., 2021) and LLMs (Touvron et al., 2023; Chiang et al., 2023). Building on top of MLLMs, some works seek to improve the quality of instruction-following data (Chen et al., 2023b; Wang et al., 2023a) or to enhance object perception by introducing control signals (Jain et al., 2023), while the others explore chain-of-thoughts idea (Kojima et al., 2022; Wei et al., 2022; Zhang et al., 2023; Zheng et al., 2024; Mitra et al., 2023). In parallel, we focus on visual representation learning and the resulting visual tables serve as inputs of MLLMs and above methods.

3 Visual Table

We introduce Visual Table, a new form of visual representation, constructed in hierarchical text. Given an image $\mathbf{I}$, a visual table $\mathbf{V}_{\mathrm{T}} = g(\mathbf{I})$ is created by a generator g. $\mathbf{V}_{\mathrm{T}}$ consists of a scene description and multiple object-centric descriptions that encompass categories, attributes, and knowledge. Thanks to structural and textual formats, visual tables support interpretability to humans and LLMs, and enable controllable table editing. More-

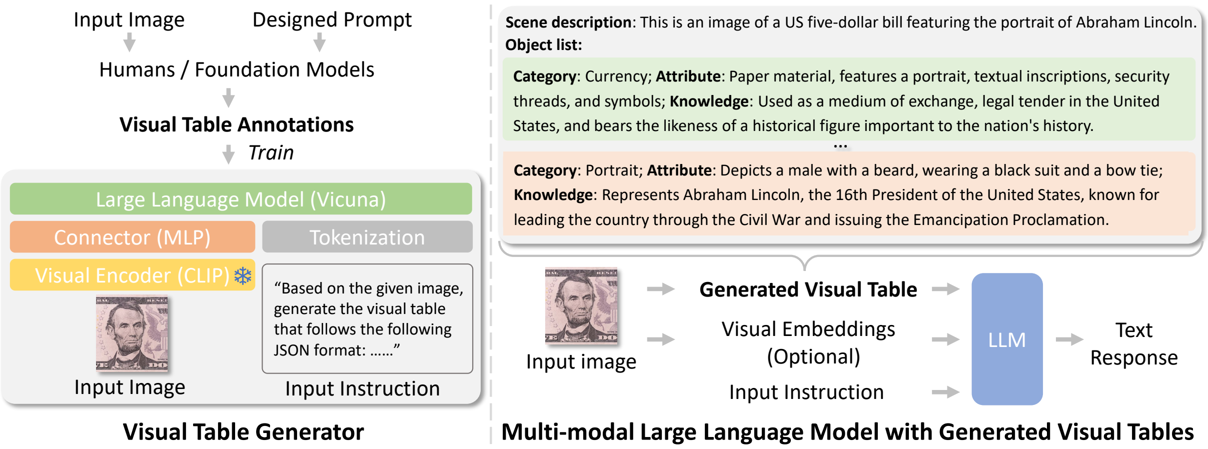


Figure 2: An overview of learning a visual table generator and its application on MLLMs. **Left:** We design a prompt to collect visual table annotations on a small scale of images. These annotations are used to train our visual table generator which consists of a frozen visual encoder, a vision-language connector, and a pre-trained LLM. **Right:** Our generator is employed to generate visual tables for the images in downstream tasks. With the generated visual table either as standalone or as additional visual representations, LLM performs reasoning to output text response.

over, they capture visual scenes at a level of granularity (*e.g.*, instance-level knowledge) and richness (*e.g.*, free-form language). These characteristics render visual tables potentially beneficial to a broad spectrum of vision tasks.

Fig. 2 provides an overview of visual table generation and its application on image reasoning using MLLMs. We first collect a small number of visual table annotations (Sec. 3.1) and then train a visual table generator g (Sec. 3.2). The trained generator is able to automatically produce visual tables for any input images in downstream tasks. We evaluate the efficacy of our generated visual tables on diverse benchmarks where visual tables serve as visual representations (Sec. 3.3).

3.1 Training Data Collection

To create visual table annotations, a straightforward solution is to ask human annotators to label images with the required visual table content. Despite high quality, training human annotators and cleaning their annotations can be expensive and cumbersome. To mitigate these burdens, we opt for an alternative solution by leveraging powerful foundation models with near-human capabilities, such as GPT4V (OpenAI, 2023). We design a detailed prompt to ensure the annotations closely align with our requirements, resulting in a small scale of visual tables, *i.e.*, on 61K images (Lin et al., 2014).

Prompt design. We design a prompt consisting of performing four tasks on input images: scene description generation, object category recognition, attribute description generation, and knowledge description generation. Without losing generality, we provide **basic definitions and guidelines** for each task to ensure versatile descriptions. For example, we define object affordance as *the functions supported by objects* and remind that *affordance might be altered case by case, due to deformed shape, unreliable materials, and so on.*

At the end of the prompt, we specify that the output should be in JSON format, *i.e.*, a nested dictionary. This structure design enables **controllable manipulation** on visual tables, such as studying the effects of removing table components. The full prompt and the collected visual tables can be found in the appendix.

Statistics. We collect a visual table per image, in total 61K images from the COCO dataset (Lin et al., 2014). These images are selected based on the scheme that an image should be associated with at least two instruction-following responses in the LLaVA-Instruct-158K dataset (Liu et al., 2023a). On average, there are 458 text tokens in each visual table after the tokenization process.

3.2 Visual Table Generation

With annotations ready, we train a model to generate visual tables. Learning such a generator is challenging and requires two key capabilities: (1) robust visual perception and (2) reliable text generation. To this end, we leverage existing MLLMs due to their remarkable performance in both visual perception and text generation.

Visual Table Generator. Our generator follows the architecture of typical MLLMs. It consists of a visual encoder that converts an input image into visual embeddings, a connector that connects visual embeddings and the LLM, and a pre-trained LLM that performs reasoning and outputs text responses. The model is trained to predict the next token in an auto-regressive way:

$$p(\mathbf{T}_{\text{a}}|\mathbf{I}, \mathbf{T}_{\text{instruct}}) = \prod_i^L p_{\boldsymbol{\theta}}(\mathbf{t}_i|h(\mathbf{I}), \mathbf{T}_{\text{instruct},<i}, \mathbf{T}_{\text{a},<i}), \quad (1)$$

where L denotes sequence length, $\boldsymbol{\theta}$ denotes trainable parameters (*e.g.*, connector and LLM), h represents frozen visual encoder, $\mathbf{T}_{\text{instruct},<i}$ and $\mathbf{T}_{\text{a},<i}$ are text tokens of instructions and answers before the current prediction token $\mathbf{t}_i$, respectively.

Training and Inference. Our generator is trained on the collected visual table annotations and partial training data from LLaVA-1.5 (Liu et al., 2023a). Specifically, there are three training stages: (1) Visual-language alignment: With visual encoder and LLM frozen, connector is trained on 595K image-text pairs by instructing the model to generate captions. (2) Instruction fine-tuning: With visual encoder frozen, connector and LLM are trained on 199K GPT-generated instruction-tuning data. (3) Supervised fine-tuning: With visual encoder frozen, connector and LLM are trained on our 61K visual table annotations. Once trained, our generator can automatically generate a visual table as the visual representation of any given image.

Note that, we avoid using any human annotations from visual question answering (VQA) datasets during training, *i.e.*, 467K VQA instances that were used to optimize VQA performance of MLLMs. This strategy aims to **minimize the biases** from VQA tasks, thereby allowing the resulting visual tables to potentially benefit a broader spectrum of visual tasks beyond VQA.

Statistics. Our generator is used to produce visual tables for all images in evaluation benchmarks, including training and test images. The average number of tokens is 421 after tokenization process.

3.3 Application on Image Reasoning

With instance-level descriptions of attributes and knowledge, visual tables essentially create comprehensive databases for individual visual scenes. In this work, we focus on its application in image reasoning tasks and assess its impact on performance within the context of MLLMs.

Specifically, visual tables are first generated for the images. We then re-train the MLLMs according to their original training methodologies, using the generated visual tables as visual representations. We choose the LLaVA-1.5 as our main testbed. The model training follows its original pipeline, except that at its second training stage (instruction-tuning stage), the LLM module takes the generated visual tables $\mathbf{V}_{\text{T}} = g(\mathbf{I})$ as the input, and learns to predict next token:

$$p(\mathbf{T}_{\text{a}}|\mathbf{I}, \mathbf{T}_{\text{instruct}}) = \prod_i^L p_{\boldsymbol{\theta}}(\mathbf{t}_i|g(\mathbf{I}), h(\mathbf{I}), \mathbf{T}_{\text{instruct},<i}, \mathbf{T}_{\text{a},<i}). \quad (2)$$

We highlight that our visual tables $g(\mathbf{I})$ are text formats and thus can function as standalone visual representations, without using visual embeddings $h(\mathbf{I})$. In this scenario, the generated visual tables can be **directly** processed by pre-trained LLMs **without** requiring the first training stage for visual-language embedding alignment. Our experiments demonstrate that using only visual tables already achieves strong performance across diverse benchmarks, while combining both visual representations leads to further performance improvements.

4 Experiments

In this section, we first introduce our implementation details, benchmarks, and evaluation protocols, and then present our results, including the comparison with traditional text-based representations, the comparison with SOTA MLLMs, the ablation study, the case study, and the human study.

Implementation Details. We provide implementation details for our visual table generator and the MLLM with visual tables. For **visual table generator**, we adopt the same architecture of LLaVA-1.5 (Liu et al., 2023a), consisting of CLIP ViT-L/14@336px (Radford et al., 2021) as the visual encoder, Vicuna-13B (Chiang et al., 2023) as the LLM, and a two-layer MLP as the connector. It is initialized by the LLaVA-1.5-13B model pre-trained by excluding VQA data, and further fine-tuned on 61K visual table annotations for 3 epochs, with batch size as 128, learning rate as 2e-5, and the optimizer as AdamW. For the **MLLM with visual table**, we adopt the same training pipeline as LLaVA-1.5, except that during the second training stage, our model is fine-tuned with the generated visual tables as additional visual representations.

Benchmarks. To evaluate visual tables, we conduct experiments across a diverse set of 11 evaluation benchmarks, providing a comprehensive assessment of visual reasoning capability. Our

Method	LLM	#PT	#IT	MM-Vet	LLaVA^{W}	MMMU	MMB	MMVP	POPE	VizWiz	SQA^{I}	GQA	VQA^{v2}	VQA^{T}
Representation: E														
BLIP-2 (Li et al., 2023c)	V-13B	129M	-	22.4	–	–	–	–	85.3	19.6	61.0	41.0	41.0	42.5
InstructBLIP (Dai et al., 2023)	V-7B	129M	1.2M	26.2	–	–	36.0	–	–	34.5	60.5	49.2	–	50.1
InstructBLIP (Dai et al., 2023)	V-13B	129M	1.2M	25.6	–	–	–	–	78.9	33.4	63.1	49.5	–	50.7
Shikra (Chen et al., 2023a)	V-13B	600K	5.5M	–	–	–	58.8	–	–	–	–	–	77.4	–
IDEFICS-9B (IDEFICS, 2023)	L-7B	353M	1M	–	–	–	48.2	–	–	35.5	–	38.4	50.9	25.9
IDEFICS-80B (IDEFICS, 2023)	L-65B	353M	1M	–	–	–	54.5	–	–	36.0	–	45.2	60.0	30.9
Qwen-VL (Bai et al., 2023)	Q-7B	1.4B	50M	–	–	–	38.2	–	–	35.2	67.1	59.3	78.8	**63.8**
Qwen-VL-Chat (Bai et al., 2023)	Q-7B	1.4B	50M	–	–	–	60.6	–	–	38.9	68.2	57.5	78.2	61.5
LLaVA-1.5 (Liu et al., 2023a)	V-7B	558K	665K	30.5	81.9	33.9	64.3	20.7	**85.9**	50.0	66.8	62.0	78.5	58.2
LLaVA-1.5 (Liu et al., 2023a)	V-13B	558K	665K	**35.4**	**86.7**	**39.5**	**67.7**	**33.3**	**85.9**	**53.6**	**71.6**	**63.3**	**80.0**	61.3
Representation: T														
Vicuna-Cap	V-13B	–	665K	23.0	79.2	39.1	62.1	12.0	73.3	51.3	69.5	48.4	61.4	48.0
Vicuna-DCap	V-13B	–	665K	27.1	77.6	37.4	61.4	13.3	**83.4**	51.6	69.3	51.7	68.5	49.6
Vicuna-SG	V-13B	–	665K	28.1	77.0	36.5	59.3	11.3	82.3	51.0	68.9	52.0	67.8	49.3
Vicuna-VT	V-13B	–	665K	**30.7**	**82.5**	**39.6**	**62.7**	**26.7**	81.9	**55.4**	**70.0**	**56.1**	**74.0**	**53.8**
Representation: E + T														
LLaVA-Cap	V-13B	558K	665K	36.3	88.9	40.6	**69.4**	32.0	86.4	53.8	71.9	63.5	80.3	60.7
LLaVA-DCap	V-13B	558K	665K	36.7	86.4	38.6	68.9	30.7	86.9	52.6	71.2	63.4	80.3	60.7
LLaVA-SG	V-13B	558K	665K	36.1	86.8	40.6	69.2	30.0	**87.3**	**57.5**	71.8	63.1	80.3	61.1
LLaVA-VT	V-13B	558K	665K	**39.8**	**89.1**	**41.9**	**69.4**	**36.7**	87.1	57.4	**72.6**	**64.0**	**80.7**	**61.2**

Table 1: **Comparison with text-based representations and MLLMs.** **E**, **T** and **E + T** denotes the visual representations as visual embeddings, text-based representations (Cap: Short Caption; Dcap: Detailed Caption; SG: Scene Graph; VT: Visual Table) and their concatenation, respectively. #PT/#IT denotes the number of samples in the stage-one/two training, respectively. V-7B/13B: Vicuna-7B/13B (Chiang et al., 2023); L-7B/13B: LLaMA-7B/13B (Touvron et al., 2023); Q-7B/13B: Qwen-7B/13B (Bai et al., 2023). Bold values refer to the best results within each group. Visual table **largely outperforms** previous text-based representations and is the only text representation that can **consistently** enhance SOTA MLLMs across diverse benchmarks.

evaluation set encompasses both recent benchmarks designed for MLLMs, including MM-Vet (Yu et al., 2023), LLaVA-Bench (Liu et al., 2023b), MMMU (Yue et al., 2023), MMBench (Liu et al., 2023c), MMVP (Tong et al., 2024) and POPE (Li et al., 2023d), and academic VQA benchmarks, including VizWiz (Gurari et al., 2018), ScienceQA (Lu et al., 2022), GQA (Hudson and Manning, 2019b), VQA-v2 (Goyal et al., 2017), and TextVQA (Singh et al., 2019). Specifically, They are diverse to cover various facets of visual reasoning. MM-Vet assesses six core vision-language capabilities, such as recognition, knowledge, OCR, spatial awareness, language generation, and math. LLaVA-Bench evaluates model capabilities on conversation, detailed description, and complex reasoning tasks that usually require world knowledge to answer accurately. MMMU evaluates models on multi-discipline tasks demanding college-level subject knowledge and deliberate reasoning. MMBench evaluates the perception and reasoning capabilities. MMVP measures visual understanding capabilities by collecting image pairs that CLIP perceives as similar despite their clear visual differences. POPE assesses the object hallucination problem. VizWiz requires models to answer questions from individuals with visual impairments. ScienceQA spans questions from subjects of natural science, language science, and social science. GQA and VQA-v2 are traditional VQA benchmarks that evaluate reasoning ability, while GQA focuses more on visual attributes. TextVQA evaluates the OCR reasoning capability.

Evaluation Protocols. We adopt two widely-used protocols: (1) Exact-matching protocol matches the predicted answer string and ground-truth string (Goyal et al., 2017; Hudson and Manning, 2019b). (2) GPT-assisted protocol relies on GPT models to measure the correctness of the predicted, open-ended answer, given the question and ground-truth answer (Yu et al., 2023; Liu et al., 2023b).

4.1 Comparison Experiments

Tab. 1 shows the results of typical text-based representations, our visual tables, and recent MLLMs.

Setup. The exact-matching evaluation is utilized for academic VQA benchmarks, including POPE, VizWiz, ScienceQA, VQA-v2, GQA, and TextVQA. This evaluation protocol is also applied to MMBench and MMVP benchmarks due to their multiple-choice settings. We use GPT-assisted evaluation for the remaining benchmarks, including MM-Vet (open-ended VQA, using the official GPT-4 evaluation server), LLaVA-Bench (open-ended VQA, using GPT-3.5-1106), and MMMU (open-ended VQA, with 855 VQA paris sampled from its original val split, using GPT-3.5-1106).

Baselines. Visual table is designed as a structural text-based representation. We thus compare it with the commonly-used text-based representations, including: (1) **Cap**: We generate short captions using BLIP2-OPT-2.7B (Li et al., 2023c), a captioner from the widely-used BLIP model family. LLaVA-1.5 also employs a captioner from this family to create its training data. (2) **DCap**: We use the same pre-trained LLaVA-1.5 model as our generator to produce detailed captions since it already learned image captioning during training. The difference is that our generator is further fine-tuned on visual table annotations. (3) **SG**: Inspired by (Mitra et al., 2023), we utilize the same pre-trained LLaVA-1.5 model as our generator to produce scene graphs, including visual attributes and visual relationships. Both scene graphs and visual tables are structural representations, while visual tables have richer semantic descriptions and cover object knowledge.

With these text-based representations ready, we replace visual tables and re-train MLLMs. When used as standalone visual representation, without the need for visual embeddings or the need for visual-language alignment training, they are denoted as **Vicuna-Cap/DCap/SG**, respectively. When combined with visual embeddings, they are denoted as **LLaVA-Cap/DCap/SG**, respectively.

Further, to validate whether visual tables can enhance existing MLLMs that have already exhibited incredible results on visual reasoning, we incorporate baselines of recent MLLMs, *e.g.*, BLIP-2 (Li et al., 2023c), InstructBLIP (Dai et al., 2023), Shikra (Chen et al., 2023a), IDEFICS (IDEFICS, 2023), Qwen-VL (Bai et al., 2023) and LLaVA-1.5 (Liu et al., 2023a).

Comparison with Text-based Representations. As Tab. 1 shows, no matter in Vicuna-VT setting or LLaVA-VT setting, visual tables **significantly** outperform traditional text-based representations (*e.g.*, +2.6 on MM-Vet, +5.5 on LLaVA-Bench, +3.1 on MMMU, +15.4 on MMVP over Vicuna-SG). Moreover, it is worth noting that visual table (LLaVA-VT) is the **only** text-based representation that can **consistently improve** the base model LLaVA-1.5 across benchmarks. These results reveal that, unlike previous text-based representations, visual tables stand out as robust visual representations capable of generalization across various scenarios.

Consistent Improvements over SOTA MLLMs. Despite the challenge of further enhancing existing MLLMs, which have already demonstrated remarkable performance in visual reasoning, visual tables (LLaVA-VT) **consistently** achieve improvements

Visual Table Scene	Visual Table Attribute	Visual Table Knowledge	GQA Accuracy	MMVP Accuracy	MM-Vet Accuracy	MMMU Accuracy
✗	✗	✗	56.7	18.0	35.6	42.3
✓	✗	✗	56.8	20.0	41.2	41.3
✓	✓	✗	**58.5**	20.0	41.8	44.0
✓	✗	✓	56.3	<u>20.7</u>	**43.7**	**45.4**
✓	✓	✓	<u>57.8</u>	**21.3**	<u>43.0</u>	<u>44.8</u>

Table 2: **Ablation study on visual table components**. We probe the components by editing visual tables and re-training our MLLM. Bold/underlined values refer to the best/second-best results.

over SOTA MLLMs across diverse benchmarks (*e.g.*, +4.4 on MM-Vet, +2.4 on LLaVA-Bench, +2.4 on MMMU, +3.4 on MMVP). These consistent enhancements across various benchmarks validate the **generality** of our visual table design and can be attributed to the valuable information provided by visual tables, such as world knowledge, which visual embeddings lack.

Textual Representations vs. Visual Embeddings. It is not surprising that visual embeddings perform better than text-based representations (LLaVA-1.5 vs. Vicuna-Cap/DCap/SG/VT). However, it is worth recognizing that text representations by nature offer **unique advantages** over visual embeddings. For instance, textual formats can be directly utilized by LLMs without the need for visual-language alignment (*e.g.*, Vicuna-VT), offering a solution for domains where alignment data is scarce, such as 3D vision. Textual formats are also interpretable to humans, facilitating applications that require human-machine interaction. Moreover, by explicitly representing object knowledge in text, visual tables can support research such as embodied AI that requires an understanding of object affordance (Gu et al., 2023). Overall, we believe that text-based representations are worth exploring and our work suggests a promising avenue of representing visual scenes in structural and rich text.

4.2 Ablation Study

Benefiting from the structural design, the components in visual tables can be effortlessly edited. This enables ablation experiments to probe the impacts of removing certain components in visual tables. We use the edited visual tables as visual representations and re-train MLLMs for comparisons.

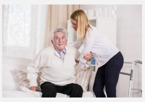
Setup. In the ablation study, we utilize 61K annotated visual tables to inspect the impacts of each component since they follow strict JSON formats. All MLLMs are first trained on 595K image-text pairs, and then fine-tuned on 177K instruction-following data with the edited visual tables as in-

Question: Please explain the meme in detail.
GT Answer: The meme features a dog lying on a cyan wooden floor, looking exhausted or unenthusiastic. … The meme humorously captures the feeling of a lack of motivation and energy that is often associated with Mondays ...

LLaVA: The image features a small dog lying on a wooden floor … The caption reads "Monday. Just Monday" … The dog appears to be relaxed and enjoying its time on the wooden surface. The scene captures a moment of leisure and contentment for the dog, as it spends its day indoors.

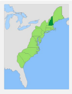
LLaVA-VT: The image features a small dog lying on a wooden floor, appearing to be sleepy or tired. The dog is positioned in the center of the scene …The meme is a play on words, with the phrase "Just Monday" written below the dog. This cleverly implies that the dog is expressing its feelings about the start of the week, as if it were a human. The meme is likely meant to evoke a sense of humor and relatability, as many people might feel tired or unenthusiastic about starting a new week.

Generated Visual Table: **Scene**: an image of a dog lying; text that says "MONDAY. JUST MONDAY."; suggesting a feeling of dread or reluctance associated with the start of the work week. **Object "Dog"**: lying down with its head resting on the floor, possibly sad; often reflect the emotions of their owners, indicating a state of tiredness. **Object "Text"**: added to express a sentiment about the start of the week; the use of text in images is a common way to convey additional messages or emotions that complement the visual content, it suggests a feeling of reluctance or negativity associated with the beginning of the work week. **Object "Wooden floor"**: ...



Question: How many people are sitting down?
GT Answer: one
LLaVA: There are two people sitting down in the image.
LLaVA-VT: There is one person sitting down, an older man.

Generated Visual Table: **Scene**: An elderly man is sitting on a couch in a home setting, smiling as a young woman stands beside him, gently assisting or comforting him. **Object "Elderly man"**: sitting, smiling, appears content; receiving care or companionship. **Object "Young woman"**: standing, assisting or comforting the elderly man; engaged in a supportive or nurturing role. **Object "Couch"**: ... **Object "Cushion"**: ... **Object "Wall decoration"**: ... **Object "Window"**: ...



Question: What is the name of the colony shown? A. Maryland B. New Hampshire C. Rhode Island D. Vermont
GT Answer: B
LLaVA: C
LLaVA-VT: B

Generated Visual Table: **Scene**: This is a map showing the eastern seaboard of the United States, with states highlighted in green. **Object "Map"**: displays the eastern seaboard of the United States, with states outlined and shaded in green, overlaid on a base of water bodies and landforms; focus on the political boundaries of the states. **Object "States"**: The states are outlined and shaded in green, indicating some sort of highlighting or grouping. The states are: Maine, New Hampshire, Vermont, Massachusetts, Rhode Island, Connecticut, New York, New Jersey, Delaware, Maryland, Virginia, North Carolina, South Carolina, Georgia, and Florida; States are the primary units of government in the United States. **Object "Water bodies"**: shown in blue and include the Atlantic Ocean, the Gulf of Mexico, and various bays, straits, and inlets; also play a role in defining state boundaries along coastlines. **Object "Landforms"**: ...

Question: Which action is performed in this image? A. pushing car B. snowboarding C. biking through snow D. shoveling snow
GT Answer: A
LLaVA: B
LLaVA-VT: A

Generated Visual Table: **Scene**: A person is engaged in a physical activity, possibly pushing or lifting a car in a snowy parking lot. **Object "Person"**: Engaged in a strenuous activity, wearing winter clothing, appears to be exerting effort; The person is likely attempting to move a vehicle that is stuck in the snow, which is a common problem during winter. **Object "Car"**: Partially covered in snow, appears to be a sedan; The car is immobile, likely due to being stuck in the snow, cars can become immovable when their tires spin on ice or snow without enough traction. **Object "Snow"**: Covers the ground; It can create hazardous conditions for vehicles if not properly managed. **Object "Parking lot"**: Covered with snow, has multiple parking spaces, some occupied by vehicles; provide a designated place for vehicles to park, they require maintenance such as snow removal to ensure accessibility. **Object "Trees"**: ... **Object "Fence"**: ... **Object "Sign"**: ...

Figure 3: Visualization of visual reasoning examples. For simplicity, we visualize partial visual tables that relate to the question, with attribute and knowledge separated by “;”.

puts. GPT-assisted evaluation is used for benchmarks: GQA (open-ended VQA with 398 VQA pairs sampled from test-dev split, using GPT-3.5-1106), MMVP (open-ended VQA, using the GPT-3.5-1106), MM-Vet (open-ended VQA, using their official GPT-4 evaluation server), and MMMU (open-ended VQA with 855 VQA paris sampled from its original val split, using GPT-3.5-1106).

Results. Tab. 2 presents the results of our ablation study. Compared to the baseline model (row 1) that takes CLIP visual embeddings as the only visual representations, scene descriptions (row 2) improve the performance on MMVP (+2.0) and MM-Vet (+5.6), yet bring limited benefits on GQA (+0.1) and worse performance on MMMU (-1.0). These results suggest that scene descriptions can provide useful information but **cannot robustly** benefit wide benchmarks. When compared with row 2, adding attributes (row 3) largely improves the performance on GQA (+1.7), and adding knowledge (row 4) significantly improves the results on MM-Vet (+2.5) and MMMU (+4.1). These results align with intuition since GQA highlights object **attributes** while MM-Vet and MMMU heavily rely on **knowledge** to answer the questions.

Combining all components, full visual tables (row 5) achieve either the best or the second-best results across all benchmarks, striking a good balance. Notably, full visual tables (row 5) largely outperforms scene descriptions (row 2), even though both are annotated by GPT4V. These results validate the necessity of **all** components and the consistent performance improvements stem from **our design** of visual tables, instead of the annotation tool.

4.3 Visualization and Case Study

Benefiting from the textual formats, visual tables allow humans to interpret and inspect what information they provide to support visual reasoning, as illustrated in Fig. 3. For simplicity, we show partial visual tables, *i.e.*, the evidence that supports the visual question answering. We highlight the wrong answers in **red**, the correct answers in **blue**, and the supporting evidence from visual tables in **green**, respectively. More examples can be found in the appendix.

Take the top-left image as an example. Our model can correctly describe the theme of the image as “sleepy or tired”, instead of the answer “relaxed and enjoying its time” with opposite semantics from the LLaVA model. Our correct answer is attributed to the generated visual table. It identifies the dog as “possibly sad”, provides **the knowledge of the dog instance** as “often reflect the emotions of their owners, indicating a state of tiredness”, and explicitly offers the **world knowledge** “the beginning of the work week” based on the text “MONDAY. JUST MONDAY”. Moreover, visual tables can provide **discipline knowledge**, such as geographic knowledge in the top-right example, “the eastern seaboard of the United States” and “The states are: Maine, New Hampshire, ...”.

Besides knowledge, visual tables can **precisely recognize object attributes**. For instance, “Elderly man: sitting; Young woman: standing, assisting or comforting the elderly man” is identified

Comparable	Annotations Win	Ours Win
35.4%	42.3%	22.3%

Table 3: **Human study** to measure the quality of our generated visual tables versus collected annotations.

in the bottom-left example, thereby facilitating the correct answer. Further, visual tables explicitly record the perceived visual objects and their knowledge, **reducing the chance of hallucinations** during LLM reasoning. Consider the bottom-right image. "Person: attempting to move a vehicle that is stuck in the snow, which is a common problem during winter" in the visual table can avoid the wrong answer of "snowboarding" which also often happens during snowy days.

In summary, while being interpretable to humans, visual tables thoroughly describe the detailed objects in visual scenes and provide precise attributes and rich knowledge, thereby consistently improving SOTA MLLMs across diverse benchmarks, as demonstrated in experiments.

4.4 Human Study

In Table 3, we conduct a human study to measure the quality of our generated visual tables versus the annotations from GPT4V. This study is done with 100 randomly sampled images from the MM-Vet dataset and 3 volunteers. The results indicate that our generated visual tables have reasonable quality and can be further improved. The detailed observations are summarized as follows: (1) They are comparable on the natural images captured in daily life. (2) Annotations are oftentimes better at recognizing the text characters (*e.g.*, tables, diagrams, equations), while 38% of our sampled images contain major texts. This indicates that the generated visual tables can be further improved by adding more text-intensive images into training data. (3) Annotations tend to output lengthy descriptions, however, some information in long descriptions is irrelevant to the image theme (*e.g.*, minor characters in corners and even wrong OCR results, extended knowledge far beyond the images, excessive speculation about the deep meaning). In contrast, our generated visual tables tend to describe factual information in images.

Discussion. We utilize the foundation model GPT4V to create visual table annotations. These annotations can alternatively be sourced from human annotators, albeit at a higher cost and with lengthy post-processing. This annotation process is designed to create general representations of visual scenes, applicable to a wide range of downstream tasks beyond VQA. Importantly, neither human annotators nor foundation models have access to any downstream task-specific information, such as the particular "question" associated with an image in the VQA task. Therefore, the improvements observed in downstream performance, as well as the distinctive structural and textual properties, are attributed to the design of the visual tables, independent of the method used to collect annotations.

5 Conclusion

In this paper, we propose **Visual Table** — a new visual representation presented in structured text. Using the collected dataset with visual table annotations, we learn a generator to produce high-quality visual tables for arbitrary input images. Beyond visual embeddings, visual tables support interpretability to humans and LLMs, enable controllable editing, and meanwhile, offer rich descriptions of attributes and knowledge for each object, thereby facilitating visual reasoning. According to extensive experiments, the resulting visual tables exhibit superior performance than previous text-based representations, and demonstrate consistent improvements over the SOTA MLLMs across diverse benchmarks. We believe our study has showcased visual tables as robust and generalizable visual representations, laying the groundwork for future research concerning visual reasoning.

Limitations. When training visual table generator, we exclude the VQA annotations to minimize the biases in VQA tasks. However, our generator might still inherit undesired biases from the training datasets and the collected annotations. Besides, visual tables require computation during generation and downstream tasks. A more efficient way to generate and utilize visual tables can be further explored in future research.

Acknowledgements. This work was supported by National Key R&D Program of China (Project No. 2022ZD0161200, 2022ZD0161201). This work is also supported by Hong Kong Research Grant Council - Early Career Scheme (Grant No. 24200223) and Hong Kong Innovation and Technology Commission Project No. ITS/228/22FP. This work was also partially funded by the Centre for Perceptual and Interactive Intelligence (CPII) Ltd under the Innovation and Technology Commission (ITC)'s InnoHK.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6077–6086. IEEE.

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *https://www.anthropic.com/claude*.

Iro Armeni, Zhi-Yang He, JunYoung Gwak, Amir R Zamir, Martin Fischer, Jitendra Malik, and Silvio Savarese. 2019. 3D scene graph: A structure for unified semantics, 3d space, and camera. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5664–5673.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023a. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*.

Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023b. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. 2023c. Can pre-trained vision and language models answer visual information-seeking questions? *arXiv preprint arXiv:2302.11713*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. 2023a. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Xingyu Fu, Sheng Zhang, Gukyeong Kwon, Pramuditha Perera, Henghui Zhu, Yuhao Zhang, Alexander Hanbo Li, William Yang Wang, Zhiguo Wang, Vittorio Castelli, Patrick Ng, Dan Roth, and Bing Xiang. 2023b. Generate then select: Open-ended visual question answering guided by world knowledge. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2333–2346, Toronto, Canada. Association for Computational Linguistics.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. 2020. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*.

Qiao Gu, Alihusein Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, et al. 2023. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. *arXiv preprint arXiv:2309.16650*.

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. 2021. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*.

Liangke Gui, Borui Wang, Qiuyuan Huang, Alex Hauptmann, Yonatan Bisk, and Jianfeng Gao. 2021. Kat: A knowledge augmented transformer for vision-and-language. *arXiv preprint arXiv:2112.08614*.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617.

Sherzod Hakimov and David Schlangen. 2023. Images in language space: Exploring the suitability of large language models for vision & language tasks. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 14196–14210, Toronto, Canada. Association for Computational Linguistics.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Roei Herzig, Alon Mendelson, Leonid Karlinsky, Assaf Arbelle, Rogerio Feris, Trevor Darrell, and Amir Globerson. 2023. Incorporating structured representations into pretrained vision & language models using scene graphs. *arXiv preprint arXiv:2305.06343*.

Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A Smith, and Jiebo Luo. 2022. Promptcap: Prompt-guided task-aware image captioning. *arXiv preprint arXiv:2211.09699*.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, et al. 2024. Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems*, 36.

Drew Hudson and Christopher D Manning. 2019a. Learning by abstraction: The neural state machine. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32. Curran Associates, Inc.

Drew A Hudson and Christopher D Manning. 2019b. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*.

Nathan Hughes, Yun Chang, and Luca Carlone. 2022. Hydra: A real-time spatial perception system for 3d scene graph construction and optimization. *arXiv preprint arXiv:2201.13360*.

IDEFICS. 2023. Introducing idefics: An open reproduction of state-of-the-art visual language model. *https://huggingface.co/blog/idefics*.

Jitesh Jain, Jianwei Yang, and Humphrey Shi. 2023. VCoder: Versatile Vision Encoders for Multimodal Large Language Models. *arXiv*.

Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. 2020. Action Genome: Actions as compositions of spatio-temporal scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10236–10247.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning (ICML)*.

Namasivayam Kalithasan, Himanshu Singh, Vishal Bindal, Arnav Tuli, Vishwajeet Agrawal, Rahul Jain, Parag Singla, and Rohan Paul. 2023. Learning neuro-symbolic programs for language guided robot manipulation. In *IEEE International Conference on Robotics and Automation*.

Mahmoud Khademi, Ziyi Yang, Felipe Frujeri, and Chenguang Zhu. 2023. MM-reasoner: A multi-modal knowledge-aware framework for knowledge-based visual question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6571–6581, Singapore. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision (IJCV)*, 123(1):32–73.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. 2023a. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023b. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023c. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Lin Li, Jun Xiao, Guikun Chen, Jian Shao, Yueting Zhuang, and Long Chen. 2024. Zero-shot visual relation detection via composite visual cues from large language models. *Advances in Neural Information Processing Systems*, 36.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. 2022. Grounded language-image pre-training. In

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023d. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 292–305.

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

Yuanze Lin, Yujia Xie, Dongdong Chen, Yichong Xu, Chenguang Zhu, and Lu Yuan. 2022. Revive: Regional visual representation matters in knowledge-based visual question answering. *Advances in Neural Information Processing Systems*, 35:10560–10571.

H. Liu and P. Singh. 2004. Conceptnet — a practical commonsense reasoning tool-kit. *BT Technology Journal*, 22(4):211–226.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. In *Advances in Neural Information Processing Systems*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2023c. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204.

Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. 2023. Compositional chain-of-thought prompting for large multimodal models. *arXiv preprint arXiv:2311.17076*.

Liliane Momeni, Mathilde Caron, Arsha Nagrani, Andrew Zisserman, and Cordelia Schmid. 2023. Verbs in action: Improving verb understanding in video-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15579–15591.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Xuran Pan, Tianzhu Ye, Dongchen Han, Shiji Song, and Gao Huang. 2022. Contrastive language-image pre-training with knowledge graphs. *Advances in Neural Information Processing Systems*, 35:22895–22910.

Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. 2023. What does a platypus look like? generating customized prompts for zero-shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15691–15701.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*.

Krishan Rana, Jesse Haviland, Sourav Garg, Jad Abou-Chakra, Ian Reid, and Niko Suenderhauf. 2023. Say-plan: Grounding large language models using 3d scene graphs for scalable task planning. In *7th Annual Conference on Robot Learning*.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, pages 146–162. Springer.

Zhenwei Shao, Zhou Yu, Meng Wang, and Jun Yu. 2023. Prompting large language models with answer heuristics for knowledge-based visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14974–14983.

Jing Shi, Yiwu Zhong, Ning Xu, Yin Li, and Chenliang Xu. 2021. A simple baseline for weakly-supervised scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16393–16402.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *ICLR*.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.

Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. 2020. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3716–3725.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Damien Teney, Lingqiao Liu, and Anton van Den Hengel. 2017. Graph-structured representations for visual question answering. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal llms. *arXiv preprint arXiv:2401.06209*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.

Johanna Wald, Helisa Dhamo, Nassir Navab, and Federico Tombari. 2020. Learning 3D semantic scene graphs from 3d indoor reconstructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Junke Wang, Lingchen Meng, Zejia Weng, Bo He, Zuxuan Wu, and Yu-Gang Jiang. 2023a. To see is to believe: Prompting gpt-4v for better visual instruction tuning. *arXiv preprint arXiv:2311.07574*.

Zhenhailong Wang, Manling Li, Ruochen Xu, Luowei Zhou, Jie Lei, Xudong Lin, Shuohang Wang, Ziyi Yang, Chenguang Zhu, Derek Hoiem, et al. 2022. Language models with image descriptors are strong few-shot video-language learners. *Advances in Neural Information Processing Systems*, 35:8483–8497.

Ziyue Wang, Chi Chen, Peng Li, and Yang Liu. 2023b. Filling the image information gap for VQA: Prompting large language models to proactively ask questions. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2874–2890, Singapore. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. 2017. Scene graph generation by iterative message passing. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 5410–5419.

An Yan, Yu Wang, Yiwu Zhong, Chengyu Dong, Zexue He, Yujie Lu, William Yang, Jingbo Shang, and Julian McAuley. 2023. Learning concise and descriptive attributes for visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Jingkang Yang, Jun CEN, Wenxuan Peng, Shuai Liu, Fangzhou Hong, Xiangtai Li, Kaiyang Zhou, Qifeng Chen, and Ziwei Liu. 2023a. 4d panoptic scene graph generation. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. 2023b. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19187–19197.

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. 2022. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3081–3089.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*.

Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3208–3216.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2023. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*.

Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. 2018. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5831–5840.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5579–5588.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.

Long Zhao, Liangzhe Yuan, Boqing Gong, Yin Cui, Florian Schroff, Ming-Hsuan Yang, Hartwig Adam, and Ting Liu. 2023. Unified visual relationship detection with vision and language models. *arXiv preprint arXiv:2303.08998*.

Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. 2024. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. *Advances in Neural Information Processing Systems*, 36.

Yiwu Zhong, Jing Shi, Jianwei Yang, Chenliang Xu, and Yin Li. 2021. Learning to generate scene graph from natural language supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1823–1834.

Yiwu Zhong, Liwei Wang, Jianshu Chen, Dong Yu, and Yin Li. 2020. Comprehensive image captioning via scene graph decomposition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 211–229. Springer.

Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. 2022. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16793–16803.

Yiwu Zhong, Licheng Yu, Yang Bai, Shangwen Li, Xueting Yan, and Yin Li. 2023. Learning procedure-aware video representation from instructional videos and their narrations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14825–14835.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Appendix

In the appendix, we provide more details in addition to our main paper: (1) additional details in annotation collection, (2) additional implementation details for our MLLMs, (3) additional results using LLMs with a smaller size, (4) additional examples of visual reasoning, and (5) additional visualization of our generated visual tables.

A Visual Table Annotation Collection

To collect visual table annotations, we create a detailed prompt for GPT4V to ensure the annotations closely align with our requirements, as illustrated in Fig. 4.

Without loss of generality, we design visual tables to cover common components in visual environments. For instance, the scene description covers time, location, and event, while object descriptions detail the attributes (*e.g.*, color, material, relationships among objects) and world knowledge (*e.g.*, a basic understanding of the physical world and social life).

B Additional Implementation Details

We provide details for response prompts of MLLMs in Tab. 4. Specifically, we follow LLaVA-1.5 (Liu et al., 2023a) to design the prompts for different benchmarks, in consideration of the settings of benchmarks. For example, we instruct the MLLM to answer a single word or phrase when the benchmark adopts exact matching evaluation (*i.e.*, VQAv2, GQA, TextVQA, POPE). Similarly, we instruct the MLLM to answer open-ended answers if the benchmark adopts GPT-assisted evaluation (*i.e.*, MM-Vet, LLaVA-Bench, MMMU), to answer the choice letter if the benchmark is evaluated in a multi-choice setting (*i.e.*, MMBench, SQA-IMG, MMVP), and to answer "Unanswerable" when the images are unable to provide sufficient information (*i.e.*, VizWiz where many images are blurred).

C Additional Experiment Results

In addition to Tab. 1 in the main paper, we show more results in Tab. 5 of in appendix. The experiment settings are the same as our main paper, including benchmarks, baselines, and evaluation protocols. The only difference is that we additionally provide results for our MLLMs in 7B size, *i.e.*, Vicuna-VT-7B and LLaVA-VT-7B.

Vicuna-VT-7B. Same as the trend in our main paper, visual tables are more effective representations than previous text-based baselines. For example, even if using an LLM with a smaller size, Vicuna-VT-7B can even outperform the baselines that utilize 13B LLMs (*e.g.*, +5.0 on LLaVA-Bench, +2.6 on MMBench, +8.0 on MMVP, +5.6 on VQA-v2 over Vicuna-SG). Again, these strong results validate our design of visual tables, providing rich knowledge and precise attributes.

LLaVA-VT-7B. Our model LLaVA-VT-7B consistently outperforms the base model LLaVA-1.5-7B across diverse benchmarks. Specifically, our model performs better not only on knowledge-intensive benchmarks (*e.g.*, +1.3 on MM-Vet, +3.1 on LLaVA-Bench, +1.1 on SQA), but also on attribute-intensive benchmarks (*e.g.*, +3.4 on MMVP, +1.1 on GQA) and general benchmarks (*e.g.*, +3.7 on MMBench, +1.3 on VQA-v2). These promising results again demonstrate that our generated visual tables work as generalizable visual representations, thereby facilitating complex visual reasoning.

D Additional Visual Reasoning Examples

With the textual formats, visual tables allow humans to interpret and inspect how they support visual reasoning, as illustrated in Fig. 5. For each image, we visualize the question, the ground-truth answer, the answer from the LLaVA-1.5-13B model, the answer of our MLLM with visual tables, and our generated visual tables. For simplicity, we show partial visual tables, *i.e.*, the evidence that supports the visual question answering. We highlight the wrong answers in **red**, the correct answers in **blue**, and the supporting evidence from visual tables in **green**, respectively.

Instance-level knowledge are provided by visual tables. For instance, the knowledge of "US five-dollar bill featuring the portrait of Abraham Lincoln" and "The Starry Night by Vincent van Gogh" is given by our visual table for the 3rd and 4th examples, respectively. Moreover, our visual table can also provide **discipline knowledge**, such as the geographic knowledge in the 5th example, "the eastern seaboard of the United States" and "The states are: Maine, New Hampshire, Vermon, ...".

Besides knowledge, visual tables can **precisely recognize the object attributes**. For instance, "Elderly man: sitting; Young woman: standing, assisting or comforting the elderly man", and "Antelope: four animals with brownish-gray fur; Giraffe:

Input Prompt

You are an AI visual assistant that can analyze a single image.
Given an image, you need to perform the task of scene description.
And then, you need to identify each object in the image.
For each object, you need to perform 3 tasks: object category recognition, attribute description generation, and knowledge description generation.

Scene description:
1. Based on the given image, please provide a short and concise description for the scene in the image, such as the location, the time of the day (e.g., morning, evening), the event, and so on.

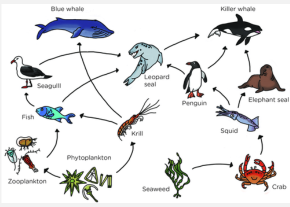
Object category recognition:
1. Based on the given image, please recognize the category for each object in the scene.
2. Please cover as many objects as possible. The objects should cover not only the salient objects, but also the other objects such as the small objects, the objects in the background, the objects that are partially occluded, and so on.

Attribute description generation:
1. Based on the given image, please generate the visual attributes for each object.
2. Visual attributes characterize the objects in images. They can be OCR characters on the object, spatial relations to surrounding objects, action relations to surrounding objects, relative size compared to surrounding objects, color, geometry shape, material, texture pattern, scene environment, motion or dynamics of objects, and so on.
3. Specially, if possible, the visual attributes could be the emotions (e.g., surprised, angry), age (e.g., young, elderly), and so on.

Knowledge description generation:
1. Based on the given image, please describe the knowledge for each object.
2. The knowledge includes object affordance, commonsense knowledge, background knowledge, and so on.
3. Object affordance is defined as the functions supported by the objects. For example, what the objects can be used for? Note that the affordance might be altered case by case, due to deformed shape, unreliable materials, and so on.
4. Commonsense knowledge is defined as basic understandings and assumptions about daily life, human behavior, and the natural world.
It also includes understanding social norms, basic cause-and-effect relationships, and simple reasoning about daily situations.
5. Background knowledge is defined as the knowledge of named entities, such as celebrities, ceremonies, festivals, and so on.

Output format:
The output content should follow the following JSON format.
{"scene description": "", "objects": [{"object category": "", "attribute description": "", "knowledge description": ""}, ..., {"object category": "", "attribute description": "", "knowledge description": ""}]}.
Directly output the JSON without any other content. The output MUST follow JSON format.

Input Image



Visual Table Annotation

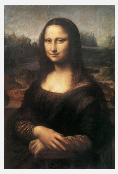
{"scene description": "The image is a simple illustration representing a marine food web, showcasing the relationships between various marine animals and plants.",
"objects": [
{"object category": "Blue whale", "attribute description": "Large, blue and gray in color, streamlined body shape", "knowledge description": "Marine mammal, known to be the largest animal on Earth, feeds on krill and other small sea creatures"},
{"object category": "Killer whale", "attribute description": "Black and white coloration, distinctive dorsal fin", "knowledge description": "Also known as orca, a toothed whale that is a top predator, feeds on various marine animals including seals and fish"},
{"object category": "Seagull", "attribute description": "White with gray wings, flying bird with a sharp beak", "knowledge description": "Seabird that feeds opportunistically, often found scavenging and eating fish or invertebrates"},
{"object category": "Leopard seal", "attribute description": "Spotted gray coat, large body with a reptile-like head", "knowledge description": "One of the top Antarctic predators, known for its diverse diet including fish, penguins, and even other seals"},
{"object category": "Penguin", "attribute description": "Black and white plumage, flightless bird with an upright stance", "knowledge description": "Adapted to marine life, feeds primarily on fish and krill and is a prey item for seals and killer whales"},
{"object category": "Elephant seal", "attribute description": "Large, thick-bodied seal with brownish skin color", "knowledge description": "Named for their large size and males' trunk-like snouts, they feed on fish and squid, and can dive deep"},
{"object category": "Fish", "attribute description": "Typical streamlined fish shape, blue with scales", "knowledge description": "General representation of fish, diverse group that can feed on plankton or be prey for larger marine animals"},
{"object category": "Krill", "attribute description": "Small, red-orange, shrimp-like creature", "knowledge description": "Crustacean vital to the marine food chain; a primary food source for many marine animals, from fish to whales"},
{"object category": "Squid", "attribute description": "Elongated body with tentacles, often depicted in brown color", "knowledge description": "Cephalopods known for their mobility, intelligence, and as a food source for predators like seals and whales"},
{"object category": "Crab", "attribute description": "Red-orange color, characterized by a round shell and pincers", "knowledge description": "Crustaceans with a varied diet; some species may feed on seaweed, mollusks, and detritus"},
{"object category": "Zooplankton", "attribute description": "Tiny, varied shapes, often transparent or with visual cues for microscopic details", "knowledge description": "Small and often microscopic organisms drifting in water, a crucial component of the marine food web"},
{"object category": "Phytoplankton", "attribute description": "Illustrated as small green organisms with simple structures", "knowledge description": "Microscopic plants or plant-like organisms in aquatic environments, they perform photosynthesis and are a base food source"},
{"object category": "Seaweed", "attribute description": "Green, plant-like with leafy fronds", "knowledge description": "Marine algae that provide habitat and food for various marine creatures, also play a role in oxygen production"}]}

Input Image

Visual Table Annotation

{"scene description": "The image depicts a pastoral scene in a dry, arid environment, likely a desert or semi-desert region during the day. A herder is tending to a flock of goats, possibly in the morning or midday judging by the shadows and the bright sky.",
"objects": [
{"object category": "Person", "attribute description": "Adult male, wearing traditional attire, standing amidst the flock, holding an object that might be a staff or tool", "knowledge description": "The person is a goat herder, responsible for the care, guidance, and protection of the livestock. The traditional attire suggests a cultural or regional dress code, and the staff or tool is typically used to assist in herding."},
{"object category": "Goats", "attribute description": "Various sizes, predominantly black and white, with long hair, some with horns, scattered across the foreground and midground of the image", "knowledge description": "Goats are domesticated animals commonly raised for their meat, milk, and wool. They are known for their hardiness and ability to graze on sparse vegetation, which makes them well-suited for this environment."},
{"object category": "Fence", "attribute description": "Wire fence in the background, enclosing an area", "knowledge description": "The fence likely serves to contain livestock or demarcate property boundaries within the pastoral area."},
{"object category": "Shack", "attribute description": "Small, rudimentary structure with a flat roof in the background", "knowledge description": "The structure may serve as storage for supplies, shelter for the herder, or a place to keep the goats during inclement weather or at night."},
{"object category": "Mountains", "attribute description": "Bare mountains visible in the far background under a clear blue sky", "knowledge description": "The mountains contribute to the geography of the region and may be part of a larger range that influences the local climate and ecosystem."}]}

Input Image



Visual Table Annotation

{"scene description": "This is an image of the Mona Lisa, a famous Renaissance painting by Leonardo da Vinci. The painting depicts a woman with an enigmatic smile, seated in front of a distant landscape.",
"objects": [
{"object category": "Painting", "attribute description": "The subject is a woman with dark hair, a subtle smile, attire from the Renaissance period, and a mountainous landscape in the background.", "knowledge description": "The Mona Lisa painting is one of the most recognized and valuable paintings in the world, housed in the Louvre Museum in Paris. It is celebrated for the subject's mysterious smile and the artist's use of sfumato, a technique that creates a soft transition between colors and tones."}]}

Figure 4: An overview of our designed prompt and the collected visual tables.

Benchmark	Response Prompts
MM-Vet, LLaVA-Bench, MMMU	–
POPE, GQA, VQAv2, TextVQA	Answer the question using a single word or phrase.
MMBench, ScienceQA-IMG, MMVP	Answer with the option's letter from the given choices directly.
VizWiz	When the provided information is insufficient, respond with 'Unanswerable'. Answer the question using a single word or phrase.

Table 4: **Response prompts** for evaluation benchmarks. The task instruction is concatenated to the input of MLLMs.

Method	LLM	#PT	#IT	MM-Vet	LLaVA^{W}	MMMU	MMB	MMVP	POPE	VizWiz	SQA^{I}	GQA	VQA^{v2}	VQA^{T}
Representation: E														
BLIP-2 (Li et al., 2023c)	V-13B	129M	-	22.4	–	–	–	–	85.3	19.6	61.0	41.0	41.0	42.5
InstructBLIP (Dai et al., 2023)	V-7B	129M	1.2M	26.2	–	–	36.0	–	–	34.5	60.5	49.2	–	50.1
InstructBLIP (Dai et al., 2023)	V-13B	129M	1.2M	25.6	–	–	–	–	78.9	33.4	63.1	49.5	–	50.7
Shikra (Chen et al., 2023a)	V-13B	600K	5.5M	–	–	–	58.8	–	–	–	–	–	77.4	–
IDEFICS-9B (IDEFICS, 2023)	L-7B	353M	1M	–	–	–	48.2	–	–	35.5	–	38.4	50.9	25.9
IDEFICS-80B (IDEFICS, 2023)	L-65B	353M	1M	–	–	–	54.5	–	–	36.0	–	45.2	60.0	30.9
Qwen-VL (Bai et al., 2023)	Q-7B	1.4B	50M	–	–	–	38.2	–	–	35.2	67.1	59.3	78.8	**63.8**
Qwen-VL-Chat (Bai et al., 2023)	Q-7B	1.4B	50M	–	–	–	60.6	–	–	38.9	68.2	57.5	78.2	61.5
LLaVA-1.5 (Liu et al., 2023a)	V-7B	558K	665K	30.5	81.9	33.9	64.3	20.7	**85.9**	50.0	66.8	62.0	78.5	58.2
LLaVA-1.5 (Liu et al., 2023a)	V-13B	558K	665K	**35.4**	**86.7**	**39.5**	**67.7**	**33.3**	**85.9**	**53.6**	**71.6**	**63.3**	**80.0**	61.3
Representation: T														
Vicuna-VT	V-7B	–	665K	28.7	82.0	33.2	61.9	19.3	81.2	52.4	67.3	55.3	73.4	51.1
Vicuna-Cap	V-13B	–	665K	23.0	79.2	39.1	62.1	12.0	73.3	51.3	69.5	48.4	61.4	48.0
Vicuna-DCap	V-13B	–	665K	27.1	77.6	37.4	61.4	13.3	**83.4**	51.6	69.3	51.7	68.5	49.6
Vicuna-SG	V-13B	–	665K	28.1	77.0	36.5	59.3	11.3	82.3	51.0	68.9	52.0	67.8	49.3
Vicuna-VT	V-13B	–	665K	**30.7**	**82.5**	**39.6**	**62.7**	**26.7**	81.9	**55.4**	**70.0**	**56.1**	**74.0**	**53.8**
Representation: E + T														
LLaVA-VT	V-7B	558K	665K	31.8	85.0	34.3	68.0	24.0	86.5	50.5	67.9	63.1	79.8	59.7
LLaVA-Cap	V-13B	558K	665K	36.3	88.9	40.6	**69.4**	32.0	86.4	53.8	71.9	63.5	80.3	60.7
LLaVA-DCap	V-13B	558K	665K	36.7	86.4	38.6	68.9	30.7	86.9	52.6	71.2	63.4	80.3	60.7
LLaVA-SG	V-13B	558K	665K	36.1	86.8	40.6	69.2	30.0	**87.3**	**57.5**	71.8	63.1	80.3	61.1
LLaVA-VT	V-13B	558K	665K	**39.8**	**89.1**	**41.9**	**69.4**	**36.7**	87.1	57.4	**72.6**	**64.0**	**80.7**	**61.2**

Table 5: **Comparison with text-based representations and MLLMs.** **E**, **T** and **E + T** denotes the visual representations as visual embeddings, text-based representations (Cap: Short Caption; Dcap: Detailed Caption; SG: Scene Graph; VT: Visual Table) and their concatenation, respectively. #PT/#IT denotes the number of samples in the stage-one/two training, respectively. V-7B/13B: Vicuna-7B/13B (Chiang et al., 2023); L-7B/13B: LLaMA-7B/13B (Touvron et al., 2023); Q-7B/13B: Qwen-7B/13B (Bai et al., 2023). Bold values refer to the best results within each group. Visual table **largely outperforms** previous text-based representations and is the only text representation that can **consistently** enhance SOTA MLLMs across diverse benchmarks.

standing in the distance, partially obscured by the terrain" are identified in the 6th and 7th examples, respectively, thereby facilitating the correct answer.

Overall, visual tables not only offer interpretability to humans but also thoroughly describe the detailed objects in visual scenes, providing precise attributes and rich knowledge. This comprehensive representation consistently enhances SOTA MLLMs across diverse benchmarks, as demonstrated in our experiments.

E Additional Examples of Generated Visual Tables

In the following tables, we visualize our generated visual tables. As these examples show, our visual tables can thoroughly encode the visual scenes in hierarchical text, by offering concise scene descriptions, accurate object categories, precise object attributes, and rich instance-level object knowledge.

Question: Please explain the meme in detail.
GT Answer: The meme features a dog lying on a cyan wooden floor, looking exhausted or unenthusiastic. The text at the top of the meme reads "MONDAY," and at the bottom, it says "JUST...MONDAY." The meme humorously captures the feeling of a lack of motivation and energy that is often associated with Mondays, represented by the dog's pose and expression.

LLaVA: The image features a small dog lying on a wooden floor, possibly a deck, with a caption above it. The caption reads "Monday. Just Monday. . ." The dog appears to be relaxed and enjoying its time on the wooden surface. The scene captures a moment of leisure and contentment for the dog, as it spends its day indoors.

LLaVA-VT: The image features a small dog lying on a wooden floor, appearing to be sleepy or tired. The dog is positioned in the center of the scene, occupying a significant portion of the image. The floor has a blue color, which adds to the overall ambiance of the scene. The meme is a play on words, with the phrase "Just Monday" written below the dog. This cleverly implies that the dog is expressing its feelings about the start of the week, as if it were a human. The meme is likely meant to evoke a sense of humor and relatability, as many people might feel tired or unenthusiastic about starting a new week.

Generated Visual Table: **Scene**: an image of a dog lying; text that says "MONDAY. JUST MONDAY."; suggesting a feeling of dread or reluctance associated with the start of the work week. **Object "Dog"**: lying down with its head resting on the floor, **possibly sad**; often reflect the emotions of their owners, indicating a state of tiredness. **Object "Text"**: added to express a sentiment about the start of the week; the use of text in images is a common way to convey additional messages or emotions that complement the visual content, it suggests a feeling of reluctance or negativity associated with the beginning of the work week. **Object "Wooden floor"**: ...

Question: Which scene category matches this image the best? A. manufactured home B. campus C. badlands D. field/cultivated
GT Answer: B
LLaVA: D
LLaVA-VT: B

Generated Visual Table: **Scene**: appears serene and is likely taken in a park or educational institution setting. **Object "Building"**: Red brick exterior, multiple windows , rectangular shape, two-story height; The building's architecture suggests it could be part of a school or institutional complex, designed for educational or administrative purposes. **Object "Sky"**: ... **Object "Tree"**: ... **Object "Lawn"**: ... **Object "Shrub"**: ...

Question: What is the face value of this banknote?
GT Answer: five <OR> 5
LLaVA: The face value of this banknote is $1.
LLaVA-VT: The face value of this banknote is five dollars.

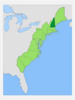
Generated Visual Table: **Scene**: an image of a US five-dollar bill featuring the portrait of Abraham Lincoln. **Object "Currency"**: Paper material, features a portrait, textual inscriptions, security threads, and symbols; Used as a medium of exchange, legal tender in the United States, and bears the likeness of a historical figure important to the nation's history. **Object "Portrait"**: ...; Abraham Lincoln, the 16th President of the United States. **Object "Text"**: Inscribed with ... "The United States of America"; Conveys the value and denomination of the currency, as well as the country of issue. **Object "Security thread"**: ...; Serves as a counterfeit deterrent feature to enhance the security of the currency. **Object "Emblem"**: An eagle and shield, symbols of the United States; Symbolizes the country's strength and heritage, commonly used in government and military insignia.

Question: Can you give a short introduction to this painting?
GT Answer: The Starry Night is an oil-on-canvas painting by the Dutch Post-Impressionist painter Vincent van Gogh. Painted in June 1889, it depicts the view from the east-facing window of his asylum room at Saint-Rémy-de-Provence, just before sunrise, with the addition of an imaginary village. It has been in the permanent collection of the Museum of Modern Art in New York City since 1941, acquired through the Lillie P. Bliss Bequest. Widely regarded as Van Gogh's magnum opus, The Starry Night is one of the most recognizable paintings in Western art.

LLaVA: The painting is a beautiful depiction of a night sky filled with stars, creating a captivating and serene atmosphere. The scene also features a mountainous landscape, adding depth and dimension to the artwork. The combination of the starry night sky and the mountainous terrain creates a sense of tranquility and wonder, inviting the viewer to appreciate the beauty of nature.

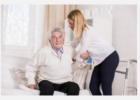
LLaVA-VT: The painting is a famous artwork by Vincent Van Gogh, titled "The Starry Night." It features a nighttime scene with a village and a church in the foreground, surrounded by a beautiful starry sky. The painting is known for its vivid colors and the artist's unique style, which captures the essence of the night and the tranquility of the village.

Generated Visual Table: **Scene**: This is a painting depicting a night scene with a starry sky. The painting is known as 'The Starry Night' by Vincent van Gogh. **Object "Starry sky"**: The sky is filled with bright stars and swirling patterns; associated with tranquility and the vastness of the universe. **Object "Moon"**: upper right of the painting; key element in the painting providing light and contrast. **Object "Tree"**: left side of the painting, reaching towards the sky; used in art to represent life and growth, as well as to add contrast and depth to a scene. **Object "Houses and buildings"**: ...; depict a rural or small town setting, providing a sense of human habitation and community. **Object "Mountain"**: visible in the background behind the village; add a sense of scale and grandeur to a landscape painting. **Object "Church tower"**: ...



Question: What is the name of the colony shown? A. Maryland B. New Hampshire C. Rhode Island D. Vermont
GT Answer: B
LLaVA: C
LLaVA-VT: B

Generated Visual Table: **Scene**: This is a map showing the eastern seaboard of the United States, with states highlighted in green. **Object "Map"**: displays the eastern seaboard of the United States, with states outlined and shaded in green, overlaid on a base of water bodies and landforms; focus on the political boundaries of the states. **Object "States"**: The states are outlined and shaded in green, indicating some sort of highlighting or grouping. The states are: Maine, New Hampshire, Vermont, Massachusetts, Rhode Island, Connecticut, New York, New Jersey, Delaware, Maryland, Virginia, North Carolina, South Carolina, Georgia, and Florida; States are the primary units of government in the United States. **Object "Water bodies"**: shown in blue and include the Atlantic Ocean, the Gulf of Mexico, and various bays, straits, and inlets; also play a role in defining state boundaries along coastlines. **Object "Landforms"**: ...



Question: How many people are sitting down?
GT Answer: one
LLaVA: There are two people sitting down in the image.
LLaVA-VT: There is one person sitting down, an older man.

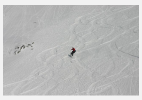
Generated Visual Table: **Scene**: An elderly man is sitting on a couch in a home setting, smiling as a young woman stands beside him, gently assisting or comforting him. **Object "Elderly man"**: sitting, smiling, appears content; receiving care or companionship. **Object "Young woman"**: standing, assisting or comforting the elderly man; engaged in a supportive or nurturing role. **Object "Couch"**: ... **Object "Cushion"**: ... **Object "Wall decoration"**: ... **Object "Window"**: ...

Question: Do these animals all have the same type?
GT Answer: No
LLaVA: Yes
LLaVA-VT: No

Generated Visual Table: **Scene**: an image of a savanna grassland during the day, with a group of antelopes in the foreground and a giraffe in the background. **Object "Antelope"**: Four animals with brownish-gray fur, standing in tall dry grass; Antelopes are herbivorous mammals known for their speed and agility. **Object "Giraffe"**: Tall animal with a long neck and legs, standing in the distance, partially obscured by the terrain; the tallest land animals. **Object "Grass"**: ... **Object "Sky"**: ...

Question: Which action is performed in this image? A. pushing car B. snowboarding C. biking through snow D. shoveling snow
GT Answer: A
LLaVA: B
LLaVA-VT: A

Generated Visual Table: **Scene**: A person is engaged in a physical activity, possibly pushing or lifting a car in a snowy parking lot. **Object "Person"**: Engaged in a strenuous activity, wearing winter clothing, appears to be exerting effort; The person is likely attempting to move a vehicle that is stuck in the snow, which is a common problem during winter. **Object "Car"**: Partially covered in snow, appears to be a sedan; The car is immobile, likely due to being stuck in the snow, cars can become immovable when their tires spin on ice or snow without enough traction. **Object "Snow"**: Covers the ground; It can create hazardous conditions for vehicles if not properly managed. **Object "Parking lot"**: Covered with snow, has multiple parking spaces, some occupied by vehicles; provide a designated place for vehicles to park, they require maintenance such as snow removal to ensure accessibility. **Object "Trees"**: ... **Object "Fence"**: ... **Object "Sign"**: ...



Question: Is there a snowboard in the image?
GT Answer: Yes
LLaVA: No
LLaVA-VT: Yes

Generated Visual Table: **Scene**: A person snowboarding down a snowy slope with visible tracks from previous skiers or snowboarders. **Object "Person"**: Wearing a red jacket, black pants, and is in a bent posture while snowboarding; engaged in snowboarding, a winter sport that involves descending a snow-covered slope while standing on a snowboard attached to the rider's feet. **Object "Snowboard"**: Partially visible under the person's feet; a piece of equipment used for snowboarding. **Object "Snow"**: Covers the ground, appears to be compacted with visible tracks; provides the necessary surface for snowboarding and skiing. **Object "Track"**: ...

Figure 5: Visualization of visual reasoning examples. For simplicity, we visualize partial visual tables that relate to the questions, with attribute and knowledge separated by ";".

Generated Visual Table	
	
User	Based on the given image, generate the visual table that follows the following JSON format: {**"scene description"**: "", "objects": [{**"object category"**: "", **"attribute description"**: "", **"knowledge description"**: ""}, ..., {**"object category"**: "", **"attribute description"**: "", **"knowledge description"**: ""}]}
Our Generator	{**"scene description"**: "This is an image of a dog lying on a wooden floor, likely inside a house. The image is accompanied by text that says ḾONDAY. JUST MONDAY.ˏ suggesting a feeling of dread or reluctance associated with the start of the work week.", "objects": [{**"object category"**: "Dog", **"attribute description"**: "The dog appears to be a small breed, possibly a French Bulldog, with a light fawn coat, lying down with its head resting on the floor, looking relaxed or possibly sad.", **"knowledge description"**: "Dogs are domesticated animals commonly kept as pets, and they often reflect the emotions of their owners. The posture of the dog suggests it is at rest, possibly indicating a state of relaxation or tiredness."}, {**"object category"**: "Wooden floor", **"attribute description"**: "The floor has a blue painted surface with visible wood grain and wear, indicating it might be an older or well-used surface.", **"knowledge description"**: "Wooden floors are a common choice for home interiors due to their durability and aesthetic appeal. The wear on the floor suggests it has been well-traveled."}, {**"object category"**: "Text", **"attribute description"**: "The text ḾONDAY. JUST MONDAY.ís superimposed on the image in a sans-serif font, likely digitally added to express a sentiment about the start of the week.", **"knowledge description"**: "The use of text in images is a common way to convey additional messages or emotions that complement the visual content. In this case, it suggests a feeling of reluctance or negativity associated with the beginning of the work week."}]}

Table 6: An example of our generated visual tables.

Generated Visual Table	
	
User	Based on the given image, generate the visual table that follows the following JSON format: {**"scene description"**: "", "objects": [{**"object category"**: "", **"attribute description"**: "", **"knowledge description"**: ""}, ..., {**"object category"**: "", **"attribute description"**: "", **"knowledge description"**: ""}]}
Our Generator	{**"scene description"**: "The image depicts an outdoor scene at dusk or dawn with a vibrant pink and orange sky. A building with a distinctive architecture is visible in the background, surrounded by lush greenery and a well-maintained lawn. The scene appears serene and is likely taken in a park or educational institution setting.", "objects": [{**"object category"**: "Building", **"attribute description"**: "Red brick exterior, multiple windows, rectangular shape, two-story height", **"knowledge description"**: "The building's architecture suggests it could be part of a school or institutional complex, designed for educational or administrative purposes."}, {**"object category"**: "Sky", **"attribute description"**: "Pink and orange hues, cloudy, dusk or dawn lighting", **"knowledge description"**: "The sky's coloration indicates that the photo was taken during the golden hour, which is shortly after sunrise or before sunset, a time prized for its soft, diffused light."}, {**"object category"**: "Tree", **"attribute description"**: "Green foliage, medium height, appears healthy", **"knowledge description"**: "Trees are commonly found in park settings and are important for providing shade, improving air quality, and enhancing the aesthetic appeal of an area."}, {**"object category"**: "Lawn", **"attribute description"**: "Green grass, well-maintained, mowed to a medium height", **"knowledge description"**: "Lawns are often found in communal outdoor areas and require regular maintenance such as mowing to keep them looking neat."}, {**"object category"**: "Shrub", **"attribute description"**: "Rounded shape, dense green leaves, medium size compared to the tree", **"knowledge description"**: "Shrubs are used in landscaping to create visually appealing boundaries and to add greenery to the environment."}]}

Table 7: An example of our generated visual tables.

Generated Visual Table	
	
User	Based on the given image, generate the visual table that follows the following JSON format: {**"scene description"**: "", "objects": [{**"object category"**: "", **"attribute description"**: "", **"knowledge description"**: ""}, ..., {**"object category"**: "", **"attribute description"**: "", **"knowledge description"**: ""}]}
Our Generator	{**"scene description"**: "This is an image of a US five-dollar bill featuring the portrait of Abraham Lincoln.", "objects": [{**"object category"**: "Currency", **"attribute description"**: "Paper material, features a portrait, textual inscriptions, security threads, and symbols", **"knowledge description"**: "Used as a medium of exchange, legal tender in the United States, and bears the likeness of a historical figure important to the nation's history."}, {**"object category"**: "Portrait", **"attribute description"**: "Depicts a male with a beard, wearing a black suit and a bow tie", **"knowledge description"**: "Represents Abraham Lincoln, the 16th President of the United States, known for leading the country through the Civil War and issuing the Emancipation Proclamation."}, {**"object category"**: "Text", **"attribute description"**: "Inscribed with 'We Do Not Do' and 'The United States of America'", **"knowledge description"**: "Conveys the value and denomination of the currency, as well as the country of issue."}, {**"object category"**: "Security thread", **"attribute description"**: "Appears as a faint, metallic strand running vertically", **"knowledge description"**: "Serves as a counterfeit deterrent feature to enhance the security of the currency."}, {**"object category"**: "Emblem", **"attribute description"**: "An eagle and shield, symbols of the United States", **"knowledge description"**: "Symbolizes the country's strength and heritage, commonly used in government and military insignia."}]}

Table 8: An example of our generated visual tables.

Generated Visual Table	
User	 Based on the given image, generate the visual table that follows the following JSON format: {**"scene description"**: "", "objects": [{**"object category"**: "", **"attribute description"**: "", **"knowledge description"**: ""}, ..., {**"object category"**: "", **"attribute description"**: "", **"knowledge description"**: ""}]}
Our Generator	{**"scene description"**: "This is a painting depicting a night scene with a starry sky. The painting is known as 'The Starry Night' by Vincent van Gogh, and it is a representation of a tranquil nighttime landscape with swirling star patterns and a bright moon.", "objects": [{**"object category"**: "Starry sky", **"attribute description"**: "The sky is filled with bright stars and swirling patterns, predominantly blue and yellow in color.", **"knowledge description"**: "The starry sky is a representation of the night sky, often associated with tranquility and the vastness of the universe."}, {**"object category"**: "Moon", **"attribute description"**: "The moon is large and bright, positioned in the upper right of the painting.", **"knowledge description"**: "The moon is a celestial body that orbits the Earth and is associated with the night; it is a key element in the painting providing light and contrast."}, {**"object category"**: "Tree", **"attribute description"**: "A tall, dark green tree stands on the left side of the painting, reaching towards the sky.", **"knowledge description"**: "Trees are often used in art to represent life and growth, as well as to add contrast and depth to a scene."}, {**"object category"**: "Houses and buildings", **"attribute description"**: "Cluster of small, dark buildings with gabled roofs, resembling a village or small town, located below the sky.", **"knowledge description"**: "The buildings are likely meant to depict a rural or small town setting, providing a sense of human habitation and community."}, {**"object category"**: "Church tower", **"attribute description"**: "A tower with a spire, illuminated, stands out against the dark sky, located in the center of the village.", **"knowledge description"**: "Church towers are often focal points in a village, serving as landmarks and symbols of community and spirituality."}, {**"object category"**: "Mountain", **"attribute description"**: "A large, dark mountain or hill is visible in the background behind the village.", **"knowledge description"**: "Mountains are natural landforms that often add a sense of scale and grandeur to a landscape painting."}]}

Table 9: An example of our generated visual tables.

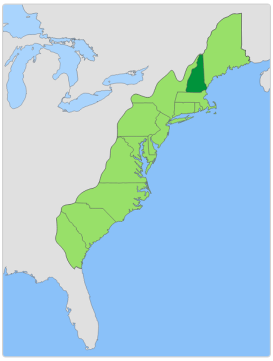
Generated Visual Table	
	
User	Based on the given image, generate the visual table that follows the following JSON format: {**"scene description"**: "", "objects": [{**"object category"**: "", **"attribute description"**: "", **"knowledge description"**: ""}, ..., {**"object category"**: "", **"attribute description"**: "", **"knowledge description"**: ""}]}
Our Generator	{**"scene description"**: "This is a map showing the eastern seaboard of the United States, with states highlighted in green. The map appears to be a political map rather than a physical or geographical map.", "objects": [{**"object category"**: "Map", **"attribute description"**: "The map displays the eastern seaboard of the United States, with states outlined and shaded in green. The map includes state names and is overlaid on a base of water bodies and landforms.", **"knowledge description"**: "Maps are used for navigation, geography education, and reference. This particular map seems to focus on the political boundaries of the states rather than physical features, which is common for political or election-related purposes."}, {**"object category"**: "States", **"attribute description"**: "The states are outlined and shaded in green, indicating some sort of highlighting or grouping. The states are: Maine, New Hampshire, Vermont, Massachusetts, Rhode Island, Connecticut, New York, New Jersey, Delaware, Maryland, Virginia, North Carolina, South Carolina, Georgia, and Florida.", **"knowledge description"**: "States are the primary units of government in the United States, each with its own set of laws, governors, and legislative bodies. The highlighting of states could be for electoral purposes, regional planning, or educational material."}, {**"object category"**: "Water bodies", **"attribute description"**: "The water bodies are shown in blue and include the Atlantic Ocean, the Gulf of Mexico, and various bays, straits, and inlets.", **"knowledge description"**: "Water bodies are essential for transportation, recreation, and are a vital part of the ecosystem. They also play a role in defining state boundaries along coastlines."}, {**"object category"**: "Landforms", **"attribute description"**: "The landforms are shown in various shades, indicating different elevations or terrain types.", **"knowledge description"**: "Landforms are natural features of the Earth's surface and are important for understanding geography, climate, and ecosystems."}]}

Table 10: An example of our generated visual tables.

Generated Visual Table	
User	 Based on the given image, generate the visual table that follows the following JSON format: {**"scene description"**: "", "objects": [{**"object category"**: "", **"attribute description"**: "", **"knowledge description"**: ""}, ..., {**"object category"**: "", **"attribute description"**: "", **"knowledge description"**: ""}]}
Our Generator	{**"scene description"**: "An elderly man is sitting on a couch in a home setting, smiling as a young woman stands beside him, gently assisting or comforting him. The environment suggests a domestic, caring interaction, possibly in a living room during the day.", "objects": [{**"object category"**: "Elderly man", **"attribute description"**: "sitting, smiling, wearing a cream-colored sweater, white hair, appears content", **"knowledge description"**: "The elderly man is likely a family member or friend, receiving care or companionship. His attire and the home setting suggest a comfortable, familiar environment."}, {**"object category"**: "Young woman", **"attribute description"**: "standing, blonde hair, wearing a white shirt and dark pants, appears to be assisting or comforting the elderly man", **"knowledge description"**: "The young woman is likely a family member, caregiver, or friend, engaged in a supportive or nurturing role. Her attire is casual, indicating an informal setting."}, {**"object category"**: "Couch", **"attribute description"**: "striped cushion, part of the furniture in the room", **"knowledge description"**: "The couch is a piece of furniture designed for seating, commonly found in living rooms and used for relaxation or socializing."}, {**"object category"**: "Cushion", **"attribute description"**: "striped pattern, placed on the couch", **"knowledge description"**: "Cushions are used for added comfort and decoration on furniture such as couches and chairs."}, {**"object category"**: "Wall decoration", **"attribute description"**: "white, appears to be a decorative item on the wall", **"knowledge description"**: "Wall decorations are used to enhance the aesthetic appeal of a room and often reflect personal tastes."}, {**"object category"**: "Window", **"attribute description"**: "partially visible, covered with a sheer curtain", **"knowledge description"**: "Windows allow natural light to enter the room and can be covered with curtains for privacy or to control the amount of light."}]}

Table 11: An example of our generated visual tables.

Generated Visual Table	
	
User	Based on the given image, generate the visual table that follows the following JSON format: {**"scene description"**: "", "objects": [{**"object category"**: "", **"attribute description"**: "", **"knowledge description"**: ""}, ..., {**"object category"**: "", **"attribute description"**: "", **"knowledge description"**: ""}]}
Our Generator	{**"scene description"**: "This is an image of a savanna grassland during the day, with a group of antelopes in the foreground and a giraffe in the background.", "objects": [{**"object category"**: "Antelope", **"attribute description"**: "Four animals with brownish-gray fur, visible horns, ears pointed upwards, standing in tall dry grass.", **"knowledge description"**: "Antelopes are herbivorous mammals known for their speed and agility. They live in herds and are commonly found in savanna ecosystems where they graze on grasses and leaves."}, {**"object category"**: "Giraffe", **"attribute description"**: "Tall animal with a long neck and legs, spotted pattern, standing in the distance, partially obscured by the terrain.", **"knowledge description"**: "Giraffes are the tallest land animals, adapted to browsing foliage higher than other herbivores can reach, usually found in African savanna environments."}, {**"object category"**: "Grass", **"attribute description"**: "Tall, dry, yellowish grass covering the ground.", **"knowledge description"**: "Grass in savannas is adapted to dry conditions and periodic fires, playing a crucial role in the ecosystem as a food source for herbivores."}, {**"object category"**: "Sky", **"attribute description"**: "Clear blue sky with no visible clouds.", **"knowledge description"**: "The clear sky suggests fair weather, which is common in savanna climates that have distinct wet and dry seasons."}]}

Table 12: An example of our generated visual tables.

Generated Visual Table	
	
User	Based on the given image, generate the visual table that follows the following JSON format: **......**
Our Generator	{**"scene description"**: "A person is engaged in a physical activity, possibly pushing or lifting a car in a snowy parking lot during the daytime.", "objects": [{**"object category"**: "Person", **"attribute description"**: "Engaged in a strenuous activity, wearing winter clothing, appears to be exerting effort", **"knowledge description"**: "The person is likely attempting to move a vehicle that is stuck in the snow, which is a common problem during winter. This requires physical strength and can be a challenging task."}, {**"object category"**: "Car", **"attribute description"**: "Partially covered in snow, stationary, appears to be a sedan", **"knowledge description"**: "The car is immobile, likely due to being stuck in the snow. Cars can become immovable when their tires spin on ice or snow without enough traction."}, {**"object category"**: "Snow", **"attribute description"**: "Covers the ground, white, appears to be compacted in areas", **"knowledge description"**: "Snow is a form of precipitation that occurs when temperatures are low enough to allow water vapor to solidify. It can create hazardous conditions for vehicles if not properly managed."}, {**"object category"**: "Trees", **"attribute description"**: "Tall, leafless, possibly dormant due to winter season", **"knowledge description"**: "Trees lose their leaves in the winter to conserve water and energy. They are often found in parking lots to provide aesthetic value and shade during warmer months."}, {**"object category"**: "Fence", **"attribute description"**: "White, wooden, appears to be a boundary marker", **"knowledge description"**: "Fences are used to mark property boundaries, provide security, and sometimes to prevent vehicles from entering certain areas."}, {**"object category"**: "Parking lot", **"attribute description"**: "Covered with snow, has multiple parking spaces, some occupied by vehicles", **"knowledge description"**: "Parking lots are common in urban and suburban areas to provide a designated place for vehicles to park. They require maintenance such as snow removal to ensure accessibility."}, {**"object category"**: "Sign", **"attribute description"**: "White, appears to be a regulatory or informational sign", **"knowledge description"**: "Signs in parking lots provide rules, directions, or information to drivers and pedestrians to maintain order and safety."}]}

Table 13: An example of our generated visual tables.

Generated Visual Table	
	
User	Based on the given image, generate the visual table that follows the following JSON format: {**"scene description"**: "", "objects": [{**"object category"**: "", **"attribute description"**: "", **"knowledge description"**: ""}, ..., {**"object category"**: "", **"attribute description"**: "", **"knowledge description"**: ""}]}
Our Generator	{**"scene description"**: "A person snowboarding down a snowy slope with visible tracks from previous skiers or snowboarders.", "objects": [{**"object category"**: "Person", **"attribute description"**: "Wearing a red jacket, black pants, and is in a bent posture while snowboarding", **"knowledge description"**: "The person is engaged in snowboarding, a winter sport that involves descending a snow-covered slope while standing on a snowboard attached to the rider's feet."}, {**"object category"**: "Snowboard", **"attribute description"**: "Partially visible under the person's feet", **"knowledge description"**: "A snowboard is a piece of equipment used for snowboarding. It is designed to glide over snow and allows the rider to maneuver through turns and tricks."}, {**"object category"**: "Snow", **"attribute description"**: "Covers the ground, appears to be compacted with visible tracks", **"knowledge description"**: "Snow is a form of precipitation that occurs when temperatures are low enough to allow water vapor to solidify. It provides the necessary surface for snowboarding and skiing."}, {**"object category"**: "Tracks", **"attribute description"**: "Curved lines in the snow made by previous skiers or snowboarders", **"knowledge description"**: "Tracks in the snow are left by the passage of skiers or snowboarders. They indicate the popularity of the slope and the skill of the individuals as tracks can vary in depth and width based on the activity and conditions."}]}

Table 14: An example of our generated visual tables.

Generated Visual Table	
	
User	Based on the given image, generate the visual table that follows the following JSON format: {**"scene description"**: "", "objects": [{**"object category"**: "", **"attribute description"**: "", **"knowledge description"**: ""}, ..., {**"object category"**: "", **"attribute description"**: "", **"knowledge description"**: ""}]}
Our Generator	{**"scene description"**: "This is an image of a food item, specifically a sandwich, against a black background with a logo in the top left corner.", "objects": [{**"object category"**: "Sandwich", **"attribute description"**: "The sandwich appears to have a variety of fillings including meats, cheese, lettuce, and tomatoes, placed between two slices of bread with visible seeds on the bread crusts.", **"knowledge description"**: "A sandwich is a common food item typically consisting of vegetables, sliced cheese or meat, placed on or between slices of bread. It is a popular choice for meals and can be found in many cultures worldwide. Sandwiches are known for their convenience as they are easy to prepare and can be customized with a variety of ingredients to cater to different tastes."}, {**"object category"**: "Logo", **"attribute description"**: "The logo is a circular emblem with a green outline and the text 'Subway Series', along with a stylized 'S' in the center. It is placed against a black background.", **"knowledge description"**: "The logo is likely associated with the Subway fast food restaurant chain, which is known for its submarine sandwiches. The term 'Series' might suggest a new promotion or a specific menu line. The 'S' could be a stylized representation of the brand's initial."}]}

Table 15: An example of our generated visual tables.

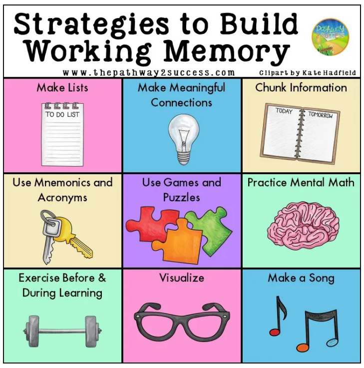
Generated Visual Table	
	
User	Based on the given image, generate the visual table that follows the following JSON format: **......**
Our Generator	{**"scene description"**: "This image is a collage of nine squares, each containing a different strategy or tip related to memory and learning, likely intended for educational purposes.", "objects": [{**"object category"**: "Square 1", **"attribute description"**: "The image shows a clip art of a pen and paper with the text 'Make lists'. The items are in black and white, and the text is in a bold, sans-serif font.", **"knowledge description"**: "The pen and paper represent writing tools commonly used for note-taking and organizing information. Making lists is a memory strategy that aids in information retention and retrieval."}, {**"object category"**: "Square 2", **"attribute description"**: "This square features a light bulb clip art with the text 'Make associations'. The light bulb is stylized with a filament inside, and the text is in a bold, serif font.", **"knowledge description"**: "The light bulb is a symbol for an idea or insight. Making associations involves connecting new information with existing knowledge to enhance memory."}, {**"object category"**: "Square 3", **"attribute description"**: "This square includes a clip art of a calendar with the text 'Review regularly'. The calendar is simple with just the days of the week visible, and the text is in a bold, serif font.", **"knowledge description"**: "A calendar is a tool for keeping track of dates and can be used for scheduling regular review sessions to reinforce learning."}, {**"object category"**: "Square 4", **"attribute description"**: "The image shows a clip art of a person with a checklist, labeled 'Visualize'. The person is depicted with a thought bubble, and the text is in a bold, serif font.", **"knowledge description"**: "Visualization involves creating mental images to aid in memory and can be used to simulate real-life scenarios for better understanding and retention."}, **......,** {**"object category"**: "Square 9", **"attribute description"**: "This square features a clip art of a person with a music note, labeled 'Make a song out of it'. The person is playing a musical instrument, and the text is in a bold, serif font.", **"knowledge description"**: "Turning information into a song can be a creative way to remember information through melody and rhythm."}]}

Table 16: An example of our generated visual tables.

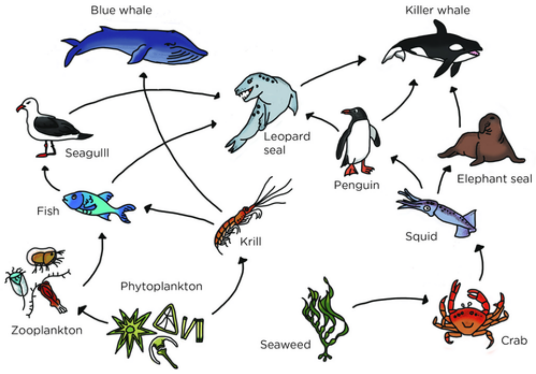
Generated Visual Table	
	
User	Based on the given image, generate the visual table that follows the following JSON format: ...
Our Generator	{**"scene description"**: "This is an illustration of a marine ecosystem with various sea creatures and plants, accompanied by arrows indicating feeding relationships and labels for each organism.", "objects": [{**"object category"**: "Blue whale", **"attribute description"**: "Large, blue, mammal, labeled 'Killer whale'", **"knowledge description"**: "The blue whale is the largest mammal in the world, known for its size and for eating other large marine animals."}, {**"object category"**: "Killer whale", **"attribute description"**: "Large, black and white, mammal, labeled 'Blue whale'", **"knowledge description"**: "Killer whales are known for their powerful hunting abilities and are often found in groups."}, {**"object category"**: "Leopard seal", **"attribute description"**: "Large, seal, spotted pattern, labeled 'Krill'", **"knowledge description"**: "Seals are semi-aquatic marine mammals that feed on a variety of prey, including krill."}, {**"object category"**: "Krill", **"attribute description"**: "Small, red, crustacean, labeled 'Leopard seal'", **"knowledge description"**: "Krill are small, shrimp-like creatures that are a staple in the diet of many marine animals."}, {**"object category"**: "Penguin", **"attribute description"**: "Medium-sized, black and white, bird, labeled 'Penguin'", **"knowledge description"**: "Penguins are flightless birds that are adapted for swimming in the ocean."}, {**"object category"**: "Seaweed", **"attribute description"**: "Green, algae, labeled 'Seaweed'", **"knowledge description"**: "Seaweed refers to various types of marine plants and algae that live in the ocean."}, {**"object category"**: "Phytoplankton", **"attribute description"**: "Microscopic, green, labeled 'Phytoplankton'", **"knowledge description"**: "Phytoplankton are tiny, photosynthetic organisms that are the base of the marine food web."}, **......,** {**"object category"**: "Sea star", **"attribute description"**: "Medium-sized, star-shaped, labeled 'Sea star'", **"knowledge description"**: "Sea stars are echinoderms with a distinctive star shape and are found in a variety of marine environments."}]}

Table 17: An example of our generated visual tables.

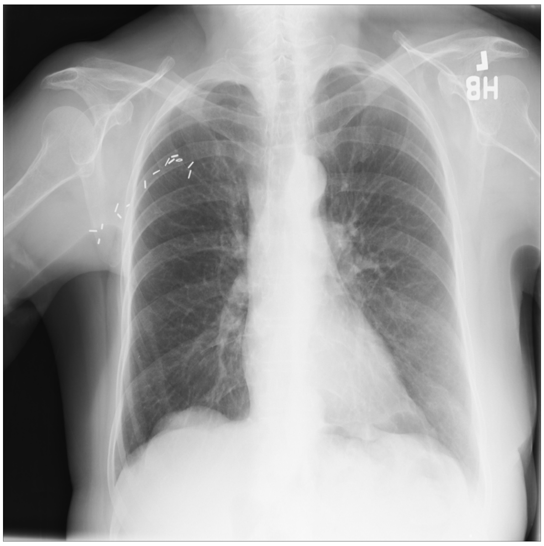
Generated Visual Table	
	
User	Based on the given image, generate the visual table that follows the following JSON format: {**"scene description"**: "", "objects": [{**"object category"**: "", **"attribute description"**: "", **"knowledge description"**: ""}, ..., {**"object category"**: "", **"attribute description"**: "", **"knowledge description"**: ""}]}
Our Generator	{**"scene description"**: "This is an X-ray image of a person's chest, highlighting the internal structures of the ribcage and the lungs.", "objects": [{**"object category"**: "Ribcage", **"attribute description"**: "Bony structure, visible as the curved skeletal framework encompassing the chest cavity, white color due to the X-ray image", **"knowledge description"**: "The ribcage provides protection for the heart, lungs, and other internal organs. It is also involved in the process of breathing, as the ribs play a role in expanding the chest to inhale air."}, {**"object category"**: "Lungs", **"attribute description"**: "Pair of spongy, air-filled organs, visible as light areas within the darker background, indicating the presence of air spaces", **"knowledge description"**: "The lungs are responsible for gas exchange, taking in oxygen and expelling carbon dioxide. They are essential for breathing and are a key part of the respiratory system."}, {**"object category"**: "X-ray label", **"attribute description"**: "Characters 'J.8H', likely indicating a code or identifier for the X-ray image, placed in the top right corner", **"knowledge description"**: "X-ray labels are used for patient identification and to help organize and retrieve medical images in a radiology department's database."}]}

Table 18: An example of our generated visual tables.

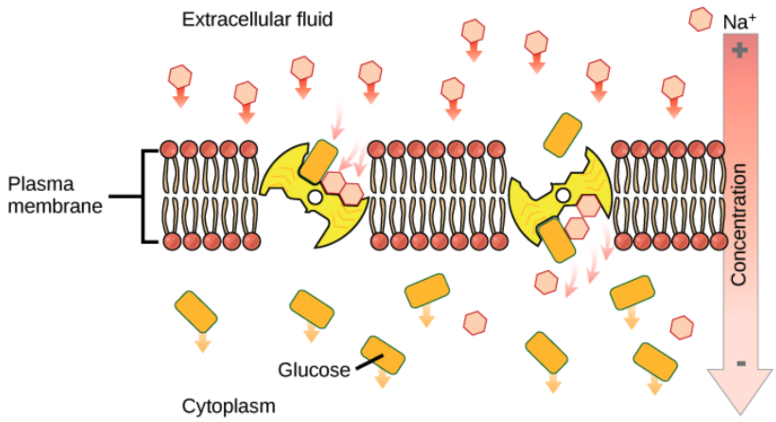
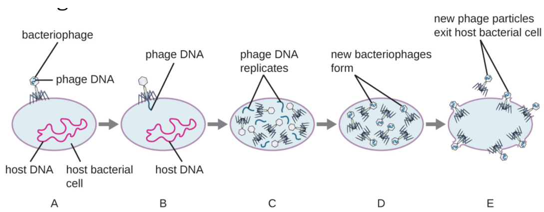
Generated Visual Table	
	
User	Based on the given image, generate the visual table that follows the following JSON format: {**"scene description"**: "", "objects": [{**"object category"**: "", **"attribute description"**: "", **"knowledge description"**: ""}, ..., {**"object category"**: "", **"attribute description"**: "", **"knowledge description"**: ""}]}
Our Generator	{**"scene description"**: "This is an illustration of extracellular fluid, specifically showing a section of connective tissue with labeled components.", "objects": [{**"object category"**: "Extracellular fluid", **"attribute description"**: "Surrounding the connective tissue cells, transparent, with label 'Extracellular fluid'", **"knowledge description"**: "The extracellular fluid is the fluid that is outside of the cells and bathes the cells, playing a crucial role in maintaining the balance of bodily fluids and transporting nutrients and waste products."}, {**"object category"**: "Connective tissue cells", **"attribute description"**: "Yellow with protrusions, labeled 'Connective tissue cells'", **"knowledge description"**: "Connective tissue cells are a diverse group of cells that provide structural support and are involved in the repair and regeneration of damaged tissues."}, {**"object category"**: "Glucose", **"attribute description"**: "Small yellow squares with arrows indicating movement, labeled 'Glucose'", **"knowledge description"**: "Glucose is a type of sugar that serves as a primary source of energy for cells and is transported through the bloodstream."}, {**"object category"**: "Cytoplasm", **"attribute description"**: "Visible within the connective tissue cells, labeled 'Cytoplasm'", **"knowledge description"**: "The cytoplasm is the gelatinous interior of a cell, where metabolic processes occur, and contains various organelles and enzymes necessary for cell function."}, {**"object category"**: "Labels", **"attribute description"**: "Text labels in English, 'Extracellular fluid', 'Connective tissue cells', 'Glucose', 'Cytoplasm'", **"knowledge description"**: "Labels are used in scientific illustrations to identify and explain the functions and components of the depicted structures."}]}

Table 19: An example of our generated visual tables.

Generated Visual Table



User

Based on the given image, generate the visual table that follows the following JSON format: **......**

Our Generator

{**"scene description"**: "This is an illustration of the process of bacterial transformation, showing a bacterial cell with foreign DNA being taken up, followed by the integration of the foreign DNA into the bacterial chromosome, and the expression of the new traits conferred by the acquired DNA.", "objects": [

{**"object category"**: "Bacterial cell", **"attribute description"**: "Ellipsoidal shape, outer membrane, cytoplasm, plasmid with replication bubbles", **"knowledge description"**: "Bacterial cells are basic units of bacteria, which are single-celled microorganisms. They have a cell wall, genetic material in form of DNA, and they replicate through a process that includes replication and transcription."},

{**"object category"**: "Plasmid", **"attribute description"**: "Small, circular DNA molecule within the bacterial cell, visible replication sites indicated by arrows", **"knowledge description"**: "Plasmids are extra-chromosomal, self-replicating genetic elements that are present in many bacteria. They can carry genes that confer antibiotic resistance or other beneficial traits for the bacteria."},

{**"object category"**: "Foreign DNA", **"attribute description"**: "Linear molecule, outside the bacterial cell, being taken up by the cell", **"knowledge description"**: "Foreign DNA can be introduced into bacterial cells through various means, such as transformation. Once inside, the DNA may integrate into the bacterial chromosome or exist as a separate plasmid."},

{**"object category"**: "Bacterial transformation", **"attribute description"**: "Uptake of foreign DNA by the bacterial cell, process indicated by arrows and a taken-up fragment of DNA", **"knowledge description"**: "Bacterial transformation is the process by which bacteria take up external genetic material, which may provide them with new traits or genes that enhance their survival."},

{**"object category"**: "Chromosome", **"attribute description"**: "Circular structure, incorporating the foreign DNA, indicating replication", **"knowledge description"**: "The bacterial chromosome is the main genetic material of the cell, where the majority of genes are located. It replicates to ensure that each daughter cell receives a copy of the genetic material."},

{**"object category"**: "New traits", **"attribute description"**: "Traits conferred by the acquired DNA, indicated by the new morphology and labeled 'New Bacterial DNA'", **"knowledge description"**: "The acquisition of new genetic material through transformation can result in the expression of novel traits in the bacteria, such as antibiotic resistance or the ability to utilize new types of substrates."}]}

Table 20: An example of our generated visual tables.

Generated Visual Table	
	
User	Based on the given image, generate the visual table that follows the following JSON format: **......**
Our Generator	{**"scene description"**: "This is an East Asian-style painting depicting a bustling scene at a bridge over a moat with people and boats. The setting appears to be in a historical or mythological context, with traditional architecture visible in the background.", "objects": [{**"object category"**: "Bridge", **"attribute description"**: "Large, arched, wooden structure with people on it", **"knowledge description"**: "Bridges are structures built to span physical obstacles without closing the way underneath such as a body of water, valley, or road, for the purpose of providing passage over the obstacle."}, {**"object category"**: "Moat", **"attribute description"**: "Wide, water-filled ditch surrounding the area", **"knowledge description"**: "Moats are often used as protective barriers around castles, cities, or other structures, and can also be part of a landscape design."}, {**"object category"**: "People", **"attribute description"**: "Various individuals depicted in traditional clothing, engaged in different activities such as walking and boating", **"knowledge description"**: "The people are likely depicted in a style that reflects the customs and societal roles of the time period or cultural setting in which the scene is meant to take place."}, {**"object category"**: "Boats", **"attribute description"**: "Small, traditional-looking boats on the water", **"knowledge description"**: "Boats are used for transportation across water and can vary greatly in size and design based on their intended use and cultural context."}, **......,** {**"object category"**: "Buildings", **"attribute description"**: "Traditional East Asian-style architecture with tiled roofs", **"knowledge description"**: "The architecture style suggests a historical or cultural significance, possibly depicting the residences or public buildings of the time."}, {**"object category"**: "Tents", **"attribute description"**: "Portable, covered shelters on the bridge", **"knowledge description"**: "Tents may be used for shelter, storage, or as market stalls, indicating commercial or communal activities."}, {**"object category"**: "Guardrails", **"attribute description"**: "Low barriers along the edges of the bridge", **"knowledge description"**: "Guardrails are safety features to prevent people and objects from falling off the edges of structures like bridges."}]}

Table 21: An example of our generated visual tables.

Generated Visual Table	
User	 Based on the given image, generate the visual table that follows the following JSON format: **......**
Our Generator	{**"scene description"**: "Two workers in protective gear are operating or maintaining machinery in an industrial setting, possibly during a work shift.", "objects": [{**"object category"**: "Person", **"attribute description"**: "Wearing a yellow hard hat, blue shirt, high-visibility safety vest, and black gloves; focused on a task; appears to be male.", **"knowledge description"**: "The attire suggests the person is a worker with a role that requires safety precautions and possibly involves manual labor or operating machinery."}, {**"object category"**: "Person", **"attribute description"**: "Wearing a yellow hard hat, blue shirt, high-visibility safety vest, and black gloves; appears to be engaged in a task; appears to be male.", **"knowledge description"**: "Similar to the first person, this individual's attire indicates a role that necessitates safety gear and manual involvement in industrial tasks."}, {**"object category"**: "Safety helmet", **"attribute description"**: "Yellow in color, worn on the head of the first person.", **"knowledge description"**: "Safety helmets are designed to protect the head from injuries due to falling objects, impacts, or other hazards."}, {**"object category"**: "Safety helmet", **"attribute description"**: "Yellow in color, worn on the head of the second person.", **"knowledge description"**: "This helmet serves the same protective purpose as the first, indicating a work environment where head protection is necessary."}, {**"object category"**: "Protective gloves", **"attribute description"**: "Black in color, worn by both individuals.", **"knowledge description"**: "Protective gloves are worn to safeguard the hands from hazards such as cuts, abrasions, or exposure to materials that could cause skin irritation or disease."}, {**"object category"**: "High-visibility safety vest", **"attribute description"**: "Bright yellow-green with reflective stripes, worn by both individuals over their shirts.", **"knowledge description"**: "High-visibility safety vests are worn to make the wearer easily visible in order to prevent accidents, especially in environments where machinery is operated or where there is vehicular traffic."}, {**"object category"**: "Machinery", **"attribute description"**: "Gray in color, with various knobs and switches, being operated or maintained by the workers.", **"knowledge description"**: "The machinery is likely part of an industrial process, and its complexity suggests it requires skilled workers to operate or maintain it properly."}]}

Table 22: An example of our generated visual tables.

Generated Visual Table	
	
User	Based on the given image, generate the visual table that follows the following JSON format: {**"scene description"**: "", "objects": [{**"object category"**: "", **"attribute description"**: "", **"knowledge description"**: ""}, ..., {**"object category"**: "", **"attribute description"**: "", **"knowledge description"**: ""}]}
Our Generator	{**"scene description"**: "A man expressing excitement or triumph with a thought bubble containing creative ideas above his head", "objects": [{**"object category"**: "Person", **"attribute description"**: "Adult male, wearing glasses, red shirt, expressing joy with mouth open and fists raised, short hair, fair skin", **"knowledge description"**: "The man's pose and facial expression suggest excitement, triumph, or a victorious moment. His gesture could be associated with success, achievement, or a creative breakthrough."}, {**"object category"**: "Thought bubble", **"attribute description"**: "Ideas depicted as light bulbs and stars, white and black, hand-drawn style", **"knowledge description"**: "Thought bubbles are a graphic convention used in art to represent what a character is thinking or dreaming. The light bulbs symbolize ideas or insights, while the stars may represent success or acclaim."}, {**"object category"**: "Light bulb", **"attribute description"**: "White, six-sided, above the person's head inside the thought bubble", **"knowledge description"**: "The light bulb is a universal symbol for an idea or a flash of inspiration. It is often used to represent creativity and the process of thinking."}, {**"object category"**: "Star", **"attribute description"**: "Yellow, five-pointed, inside the thought bubble", **"knowledge description"**: "Stars are often used to signify excellence, success, or importance. In this context, it may be depicting a sense of achievement or celebration of an idea."}]}

Table 23: An example of our generated visual tables.