UnifiedVisual: A Framework for Constructing Unified Vision-Language Datasets

Anonymous ACL submission

Abstract

Unified vision large language models (VLLMs) have shown remarkable progress in both multimodal understanding and generation, enabling tasks such as visual question answering and image generation. However, existing datasets often fall short of fully leveraging the synergistic potential between these two capabilities, thereby limiting the performance of unified VLLMs. To address this gap, we propose a novel dataset construction framework, **UnifiedVisual**, and introduce **UnifiedVisualData**, a high-quality dataset designed to enhance the mutual reinforcement between multimodal understanding and generation. UnifiedVisualData integrates both visual and textual inputs and outputs, fostering holistic multimodal reasoning and precise text-guided image generation. Moreover, the dataset demonstrates significant diversity in tasks and data sources, effectively addressing key limitations of existing datasets. To validate the effectiveness of UnifiedVisualData, we trained a unified VLLM, Anole-UnifiedVisual, which consistently outperforms models trained on existing datasets across a wide range of tasks. Notably, our model exhibits significant mutual enhancement between multimodal understanding and generation, underscoring the advantages of our framework. We believe UnifiedVisual represents a new growth point for advancing unified VLLMs and unlocking their full potential. [1]

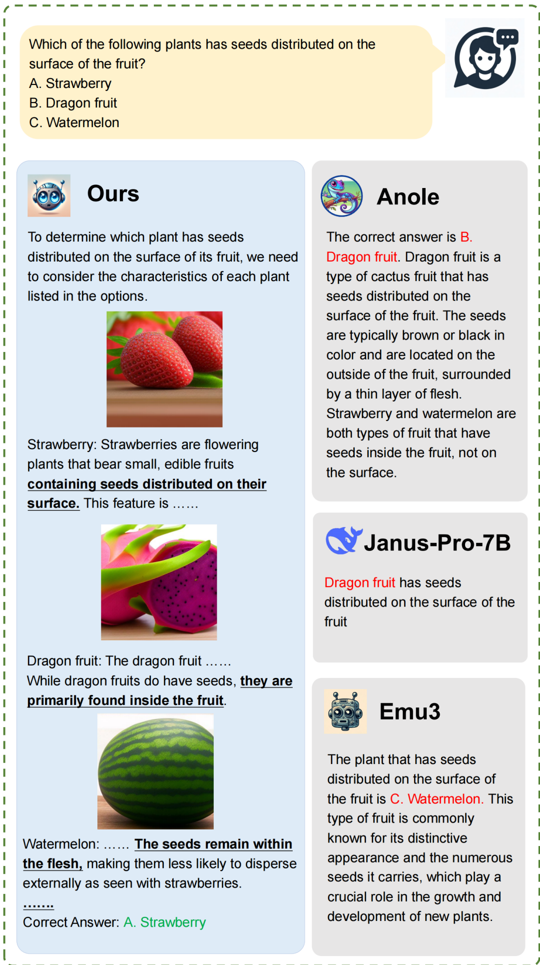


Figure 1: We introduce Anole-UnifiedVisual, a model trained on **UnifiedVisualData**, which that demonstrates outstanding mutual enhancement between multimodal understanding and generation. As illustrated in this figure, the model excels at reasoning by constructing a comprehensive multimodal reasoning chain.

1 Introduction

Vision large language models (VLLMs) have made significant progress in visual understanding, evolving from basic image captioning to complex visual inferences (Liu et al., 2024b; Dai et al., 2023). Currently, there is growing interest in unified models capable of both multimodal understanding and generation. These models aim to integrate multimodal understanding and generation capabilities, enabling them to handle a variety of tasks such as image captioning, visual question answering, and image generation (Team, 2024; Wu et al., 2024; Tong et al., 2024a). Unified VLLMs have gained widespread attention due to their ability to combine multimodal understanding and generation in a single model. This unification not only simplifies

[1]Our code and datasets will be available at https://github.com/.

the deployment and application process but also provides the potential for mutual enhancement between generative and discriminative capabilities. As a result, this area of research is becoming an increasingly prominent development field.

However, the development of powerful unified VLLMs hinges on access to high-quality training datasets. While several existing datasets have facilitated progress, they fall short of fully unlocking the synergistic potential between multimodal understanding and generation. Ideally, a unified VLLM should achieve substantial improvements by leveraging the interaction between these two capabilities. Yet, in practice, models trained on current datasets often exhibit limited integration, failing to achieve effective mutual reinforcement between understanding and generation (Wang et al., 2024b,a). This highlights a critical limitation in the design and quality of existing datasets, which are unable to fully stimulate the desired synergy.

To address these challenges, we propose a novel dataset construction framework, **UnifiedVisual**, and introduce **UnifiedVisualData**, a new dataset designed to enhance the interaction between multimodal understanding and generation. UnifiedVisualData incorporates the following key features: First, the instructions may include both visual and textual information, encouraging holistic integration of multimodal context for accurate responses. Second, the responses may also consist of both visual and textual elements, requiring the model to excel in both textual reasoning and multimodal generation. This duality ensures that textual reasoning guides precise image generation, while the generated images, in turn, enhance textual reasoning. This mutual reinforcement between the two modalities enables the model to achieve superior performance. Finally, UnifiedVisualData exhibits significant diversity in both task types and data sources, effectively promoting the interaction between understanding and generative capabilities.

To validate the effectiveness of UnifiedVisualData, we trained a unified VLLM model, Anole-UnifiedVisual, using this dataset. Experimental results demonstrate that our model consistently outperforms those trained on existing datasets across a wide range of tasks. Notably, we observed significant mutual enhancement between the model's understanding and generative capabilities, fully showcasing the advantages of our dataset.

In summary, our contributions are as follows:

- We propose UnifiedVisual, a unified vision-language dataset construction framework that prioritizes the synergistic interaction between understanding and generative capabilities while ensuring task and data source diversity.
- We construct UnifiedVisualData, a high-quality dataset tailored for unified VLLMs.
- Experimental results demonstrate that models trained on UnifiedVisualData achieve superior performance and exhibit mutual enhancement between multimodal understanding and generation.

2 Related Work

Unified Visual Understanding and Generation. In recent years, research on unifying image understanding and generation within a single visual large language model (VLLM) has garnered significant attention. Early studies primarily achieved image generation by integrating image generation models (e.g., diffusion models) on top of large language models (LLMs) (Sun et al., 2023; Wu et al., 2023; Li et al., 2024c; Ge et al., 2024). More recently, Tong et al. (2024a) demonstrated remarkable results by connecting LLMs and diffusion models through a simple projection layer. Inspired by the success of LLMs in next-step prediction tasks, recent studies have explored representing and generating images in a fully autoregressive manner using discrete visual tokens (Yu et al., 2023; Chen et al., 2023; Wang et al., 2024b; Liu et al., 2024a; Chern et al., 2024). To achieve high performance in both image understanding and generation, some research efforts have proposed decoupling these two tasks. For instance, Transfusion (Zhou et al., 2024) and Show-o (Xie et al., 2024) employ autoregressive text modeling for image understanding tasks while adopting visual diffusion modeling to accomplish image generation. In contrast, Janus (Wu et al., 2024) introduces two distinct image representations, specifically designed to address the differing granularity requirements of image understanding and generation. Overall, exploration of unified VLLM architectures continues to progress.

In this study, we evaluate performance on our dataset using Anole (Chern et al., 2024), a model built upon Chameleon (Team, 2024) that leverages a VQ Tokenizer to encode images. Anole features a unified training and testing framework and is highly representative. Compared to models such as Janus and Show-o, Anole is better suited for tasks requiring long, multimodal content in outputs.

Consequently, all experiments and analyses in this paper are conducted using Anole.

Training Datasets for Unified VLLM. Given the unique characteristics of unified VLLMs, we divided the training dataset into four major categories, as shown in Figure 2. Among them, datasets aimed at generating pure text are not only abundant but also of high quality (Shao et al., 2024b; Li et al., 2024a; Zhang et al., 2024). In contrast, datasets for multimodal generation are relatively narrow in scope and limited in scale. Currently, the most widely used multimodal generation datasets mainly include image generation datasets and image editing datasets (Deng et al., 2009; Brooks et al., 2023; Fu et al., 2023; Qu et al., 2024). Additionally, there exist interleaved image-text datasets crawled from the internet, but the association between images and text in such datasets is often weak (Zhu et al., 2024; Laurençon et al., 2024).

The scarcity of multimodal generation datasets not only limits the application of models in related downstream tasks but also introduces potential conflicts between multimodal understanding and generation during training. These conflicts make it challenging to achieve mutual enhancement of the two capabilities, potentially impacting the model's performance on complex tasks. To address these challenges, we propose a unified vision-language dataset construction framework to overcome the current limitations in training datasets.

3 Methodology

In this section, we first provide a detailed introduction to our vision-language dataset construction framework, **UnifiedVisual**. Following that, we introduce **UnifiedVisualData**, a dataset constructed following the UnifiedVisual framework.

3.1 UnifiedVisual

As discussed in Section 2, the training datasets for unified VLLM can be categorized into two types: **understanding datasets** that only contain pure text outputs, and **generation datasets** that involve multimodal generation. Given that existing understanding datasets are not only abundant but also of high quality, we can directly select from these established resources. In contrast, generation datasets tend to be relatively narrow in scope and limited in scale. To address this, UnifiedVisual introduces a novel and comprehensive framework for constructing generation datasets. Specifically, we focus on

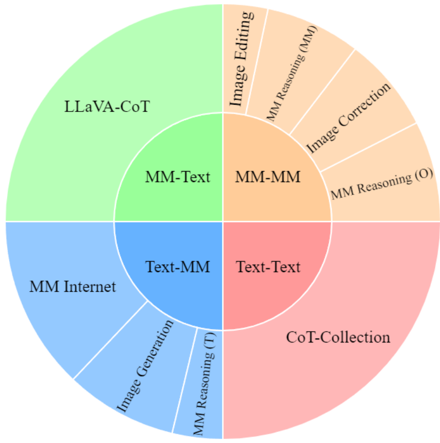


Figure 2: The proportions of different sub-datasets in UnifiedVisualData. The innermost layer of the chart represents the "*input type - output type*", such as Text-MM, which indicates that these datasets feature textual input and multimodal output.

three key aspects to construct a more diverse and comprehensive generation dataset: (1) **Visual Generation**, (2) **Multimodal Reasoning**, and (3) **Multimodal Internet Dataset**. In the sections that follow, we will discuss each construction method in detail. **The complete prompts** can be found in Appendix A.

3.1.1 Visual Generation

Visual Generation encompasses Image Generation, Image Editing, and Image Correction. Unlike existing datasets that primarily focus on generating or editing images based on simple descriptions or instructions, our goal is to integrate visual understanding and textual reasoning to tackle more complex visual generation challenges.

Image Generation Image generation involves generating images that correspond to textual descriptions, serving as a foundational task in training unified VLLMs. However, existing image generation datasets often emphasize direct mappings between textual elements and images, which limits their ability to handle more intricate generation requirements. To address this, we propose two enhanced approaches:

Topic- and Scene-Based Generation: (1) We propose several topics and corresponding scenes, then generate image captions that implicitly, rather than explicitly, describe the desired image content. (2) We use embedding models to filter duplicate captions, ensuring data diversity. (3) We use GPT-4 to generate a reasoning process (rationale) explaining the content and details to be generated, followed by DALL-E-3 for image synthesis.

Category- and Image-Based Generation: (1) We collect a diverse set of authentic images, removing duplicates. (2) Based on the images and their categories, we use GPT-4 to generate instructions that describe image characteristics and related information, again focusing on implicit rather than explicit descriptions. (3) We then use GPT-4 to generate a detailed rationale based on the caption and category information, outlining the logic behind the desired image. The final data point consists of the caption, rationale, and the original image.

Image Editing Existing image editing datasets typically consist of simple pairs of images and editing instructions that require straightforward modifications. However, these basic datasets may not effectively enhance a model's capacity to comprehend and execute sophisticated visual generation instructions. To address this limitation, we enhance existing image editing data through a two-step approach: (1) We transform simple editing instructions into more nuanced prompts that necessitate deeper understanding and planning. (2) We use GPT-4o to analyze these enhanced instructions and generate reasoning rationales outlining the editing objectives and intended outcomes.

Image Correction To further enhance the model's capability in capturing fine-grained image details, we introduced a more sophisticated task paradigm: image correction. This task requires the model to evaluate image-description consistency and, when discrepancies are identified, analyze the inconsistencies before regenerating an image that fully aligns with the given description. We implement this through a three-stage process: (1) We modify existing image captions to create descriptions that maintain the core theme while introducing controlled variations in specific visual elements. (2) We utilize StableDiffusion to generate images containing intentional discrepancies based on these modified descriptions. (3) We employ GPT-4o to systematically analyze the generated images against the original descriptions, automatically identifying inconsistencies and providing detailed modification rationales. The final data point includes the original caption, the generated image, the analysis rationale, and the original image.

3.1.2 Multimodal Reasoning

Multimodal Reasoning focuses on the synergistic interplay between multimodal understanding and generation. During the reasoning process, multimodal reasoning drives the generation of necessary visual content, while the generated visual content, as part of the reasoning rationale, in turn facilitates better multimodal understanding. This design emulates human reasoning processes, where individuals often combine textual thinking with visual aids (such as mental sketches or imagined scenes) to collaboratively solve complex problems.

MM Reasoning (O) In multimodal tasks, answering questions often requires careful attention to specific details within the **original input images**. Following Shao et al. (2024a), we construct questions that demand reasoning rationale incorporating snapshots of critical details from the original image.

MM Reasoning (MM) To enhance the model's multimodal reasoning capabilities, we construct data points that require joint reasoning across both image and text modalities. The dataset construction process is as follows: (1) We collect a diverse set of images and use the CLIP model to remove duplicates. (2) GPT-4o is employed to generate reasoning questions based on the collected images. These questions are designed to require reasoning processes that integrate both visual and textual content. Questions that fail to meet this criterion are discarded. (3) The input image and the generated question are provided to GPT-4o, which produces a rationale. When necessary, textual descriptions are used in place of images. (4) The textual descriptions from step 3 are rewritten into keywords using GPT-4. These keywords are then used to retrieve images from tools like Bing Search, ensuring stylistic consistency across the images used in the questions and rationales. (5) CLIP similarity scores are computed between the descriptions generated in step 3 and the retrieved images. Only the images with the highest similarity scores are retained. Because **the input is multimodal**, we refer to this construction method as MM Reasoning (MM).

MM Reasoning (T) Beyond multimodal input scenarios, we also design multimodal reasoning task based on **purely textual inputs**. The process is as follows: (1) GPT-4 is used to generate text-only questions that require reasoning aided by generated images. (2) The generated questions are deduplicated using embedding model to ensure diversity and uniqueness. (3) GPT-4 is then tasked with answering these questions, providing detailed rationales while replacing image-related components with textual descriptions when necessary. (4) DALL-E-3 is used to generate images based on the image descriptions generated in step 3.

3.1.3 Multimodal Internet Dataset

To further enhance the diversity and naturalness of the dataset, we process and transform interleaved text-image data sourced from the internet.

MM Internet We construct this dataset based on a large collection of diverse multimodal data crawled from the internet. To improve data quality, we draw inspiration from Chen et al. (2024b) and design a multi-perspective filtering strategy. This strategy leverages pre-trained VLLMs to ensure coherence and semantic consistency between sentences and their associated images. Furthermore, we generate questions for these multimodal data, ensuring that the answers align precisely with the corresponding text-image data.

3.2 UnifiedVisualData

Using the above methods, we ultimately constructed 120k **generation samples**. The sources and final quantities of each type of data are shown in Table 1. Additionally, we sampled 60K data points from LLaVA-CoT (Xu et al., 2024) and CoT-Collection (Kim et al., 2023), respectively, to create our **understanding samples**. Together with the generation samples, these form our UnifiedVisualData. Its composition and distribution are illustrated in Figure 2. More details about the dataset construction can be found in Appendix B.

	Quantity	Source
MM Internet	29,399	CoMM (Chen et al., 2024b)
Image Editing	9,024	MagicBrush (Zhang et al., 2023)
Image Generation	22,755	OpenImages (Krasin et al., 2017)
Image Correction	20,000	ShareGPT4V (Chen et al., 2024a)
MM Reasoning (O)	21,000	Visual-CoT (Shao et al., 2024a)
MM Reasoning (T)	7,276	-
MM Reasoning (MM)	17,761	COCO (Lin et al., 2014)

Table 1: The quantities and sources of each type of generation data in UnifiedVisualData are presented. Here, "sources" refer to the raw data sources used to construct UnifiedVisualData.

4 Experimental Setup

4.1 Unified VLLM

Architecture As analyzed in Section 2, we select Anole as the foundation for training and evaluation. Among all open-source unified VLLMs, Anole stands out as a representative model built on the transformer architecture. It adopts a unified processing approach for various modalities and supports multimodal outputs that can include any number of images. These capabilities make Anole particularly suitable as the base model for our experiments. Specifically, Anole represents images as discrete tokens. After generating these image tokens, the image decoder converts the discrete visual tokens back into images.

Training Procedure Since both the input and output may simultaneously contain text and image content, markers [BOI] and [EOI] are added before and after the visual tokens generated from the discretization of each image. With visual signals fully converted into discrete tokens, we use the standard cross-entropy loss to train the model on the next-token prediction task. Particularly, to mitigate conflicts between visual and text generation during training, we compute the loss only for text tokens when predicting text, ignoring the logits of multimodal tokens. Similarly, during visual generation, we compute the loss only for visual tokens.

Inference During inference, our model employs the next-token prediction approach. When generating text tokens, the model considers only text tokens. Once [BOI] is predicted, it signals the generation of an image. At this stage, the model focuses exclusively on predicting visual tokens until the image generation is complete.

4.2 Evaluation and Metrics

Multimodal Understanding To evaluate multimodal understanding capabilities, we conduct evaluations on six widely-used benchmarks: RealworldQA (XAI, 2024), MMVP (Tong et al., 2024b), ScienceQA (Lu et al., 2022), VStar (Wu and Xie, 2023), MME (Fu et al., 2024), and POPE (Li et al., 2023b). For RealworldQA, MMVP, ScienceQA, and VStar, accuracy is used as the evaluation metric. GPT-4 is employed to determine whether the model's output match the ground truth, and accuracy is then calculated. Notably, for MMVP, a response is only considered correct if both paired questions are answered correctly. For MME and POPE, we first use GPT-4 to summarize the model's output as either "yes" or "no" and then use the official repository's code to compute the final metrics. Specifically, for MME, we report the total score for MME Perception and MME Cognition. For POPE, we report its F1 score.

Model	RWQA	MMVP	SQA	VStar	MME	POPE	Avg.
Anole	32.0	10.0	46.7	15.7	841.4	65.8	33.4
Anole-NormalData	<u>37.9</u>	7.3	53.4	<u>30.9</u>	952.9	<u>75.9</u>	39.9
Anole-UnifiedVisual$_T$	<u>37.9</u>	<u>20.0</u>	55.2	29.8	<u>1316.5</u>	72.1	<u>43.7</u>
Anole-UnifiedVisual$_{MM}$	36.1	14.7	<u>55.3</u>	28.3	1125.3	70.6	40.9
Anole-UnifiedVisual	**39.7**	**24.0**	**56.2**	**33.0**	**1371.2**	**76.1**	**46.3**

Table 2: This table presents the results of the multimodal understanding evaluation. The best results are highlighted in **bold**, while the second-best results are marked with an <u>underline</u> for clarity.

Multimodal Generation To evaluate visual generation capabilities, we use the GenEval (Ghosh et al., 2024) benchmark. GenEval is a challenging text-to-image generation benchmark designed to reflect comprehensive generative abilities. We use the official evaluation code[2] for assessment and report the overall score.

Textual Reasoning To assess the model's pure text reasoning ability, we use AlpacaEval (Li et al., 2023a). Following the official AlpacaEval[3], we use GPT-4 for evaluation. A higher win rate indicates greater helpfulness of the response.

4.3 Experimental Details

During training, we utilized 64 NVIDIA H100 80G GPUs, set the batch size to 512, and the maximum sequence length to 4096. We used the AdamW optimizer with a 5% warm-up step and the cosine decay learning rate scheduler. The model was trained for 2 epochs with a maximum learning rate of 2e-5.

5 Experiments

5.1 Baselines

Anole-NormalData Following prior works (Ma et al., 2024; Li et al., 2024b), we trained Anole using a combination of textual understanding data, multimodal understanding data, and multimodal generation data. Specifically, the understanding data is identical to that of UnifiedVisualData, while the multimodal generation data was derived from an equivalent amount of Laion[4] (Schuhmann et al., 2022). Laion is a high-quality dataset carefully filtered by high aesthetic scores, making it a popular choice for training advanced image generation models (Xie et al., 2024). This data was subsequently transformed into the instruction-following format as outlined by Tong et al. (2024a).

[2] https://github.com/djghosh13/geneval
[3] https://github.com/tatsu-lab/alpaca_eval
[4] https://huggingface.co/datasets/dclure/laion-aesthetics-12m-umap

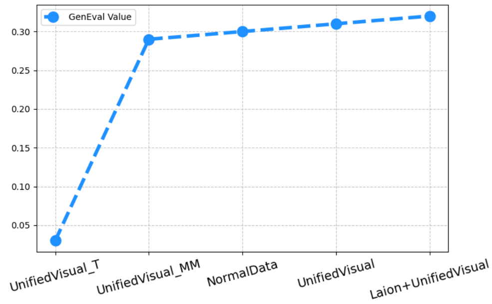


Figure 3: GenEval scores of different models.

Anole-UnifiedVisual$_T$ To investigate the interaction between multimodal understanding and generation within UnifiedVisualData, we introduced an additional baseline model trained exclusively on the understanding subset of UnifiedVisualData.

Anole-UnifiedVisual$_{MM}$ Similarly, we added another baseline model trained solely on the generation subset of UnifiedVisualData.

5.2 Main Results

5.2.1 Multimodal Understanding

The experimental results are presented in Table 2. As shown, compared to Anole-UnifiedVisual$_T$, which is trained solely on multimodal understanding data, Anole-NormalData incorporates additional multimodal generation data during training. However, its performance is notably worse than Anole-UnifiedVisual$_T$. This observation aligns with findings from prior research (Wang et al., 2024b), which indicate that directly including multimodal generation data can conflict with the training objectives of multimodal understanding tasks, leading to a decline in performance compared to training exclusively on understanding data.

In contrast, our generation data is designed not only to enhance the model's generative capabilities but also to integrate complex rationales into generation tasks. Consequently, even Anole-UnifiedVisual$_{MM}$, which is trained exclusively on

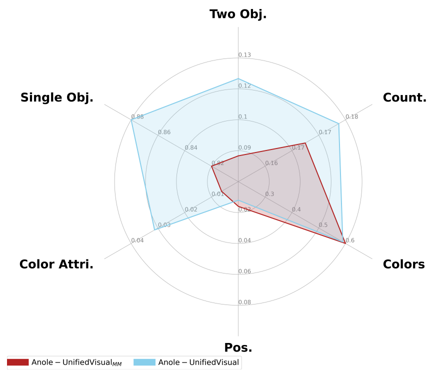


Figure 4: GenEval scores across distinct dimensions.

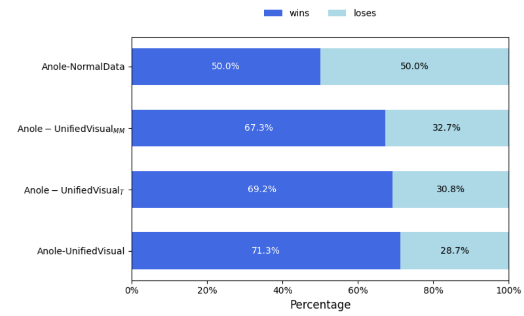


Figure 5: Evaluation on AlpacaEval.

our generation data, achieves slightly better performance than Anole-NormalData. More importantly, trained on the combined understanding and generation data we constructed, Anole-UnifiedVisual's understanding capability surpasses both Anole-UnifiedVisual$_T$, which is trained on our understanding data, and Anole-UnifiedVisual$_{MM}$, which is trained on our generation data.

These results clearly demonstrate that the generation data and understanding data in UnifiedVisualData are mutually beneficial, jointly enhancing the multimodal understanding capability of Anole-UnifiedVisual.

5.2.2 Multimodal Generation

As shown in figure 3, when trained solely on understanding data, Anole-UnifiedVisual$_T$ exhibits significantly worse generation capabilities compared to Anole-NormalData. The performance of Anole-UnifiedVisual$_{MM}$, trained on our generation data, is also slightly worse than Anole-NormalData, which may be due to the lower image quality in our UnifiedVisual dataset compared to Laion. However, when training on both the understanding and generation data in UnifiedVisualData, the generation capability of Anole-UnifiedVisual surpasses that of Anole-NormalData. This demonstrates that in UnifiedVisual, multimodal understanding data and multimodal generation data indeed promote each other, jointly enhancing the model's multimodal generation capability.

We further analyzed the detailed metrics of GenEval, as shown in figure 4. Compared to Anole-UnifiedVisual$_{MM}$, which was trained solely on generation data, Anole-UnifiedVisual achieves significant improvements in single/double-object generation, color, and quantity. This indicates that incorporating multimodal understanding data enhances the model's comprehension of object details, including attributes such as color and quantity, thereby improving its generation capability.

To further demonstrate the advantages of the generation data in UnifiedVisual over that in NormalData, we mixed half of the NormalData generation data with half of the UnifiedVisual generation data, while keeping the understanding data consistent, and trained a new model. The resulting model achieved further improvements in generation capabilities. **Compared to** Anole-UnifiedVisual, this mixed-data model benefited from the introduction of higher-quality image generation data (from Laion), leading to enhanced generation performance. This finding highlights that improving image quality can further boost model performance. Additionally, **compared to** Anole-NormalData, the introduction of more complex reasoning-based generation tasks and multimodal reasoning tasks significantly enhanced the model's generation capabilities. This further demonstrates the effectiveness of our UnifiedVisual Framework.

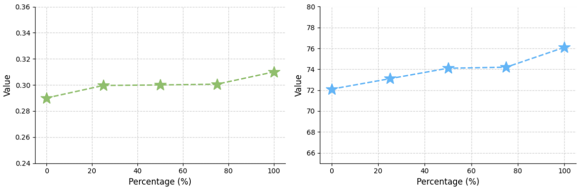
5.2.3 Text Understanding

We used AlpacaEval to evaluate the models' text understanding and problem-solving capabilities. As shown in figure 5, we calculated the win rate of all models compared to Anole-NormalData. Similar to the evaluation results for multimodal understanding, Anole-NormalData performs the worst, while Anole-UnifiedVisual achieves the best results. This once again demonstrates that in UnifiedVisualData, generation data and reasoning data mutually promote each other, thereby enhancing the model's (textual) understanding capability.

6 Analysis

6.1 Ablation study

In this section, we further demonstrate that training the model on UnifiedVisualData reveals a mutually



(a) Generation (b) Understanding

Figure 6: **Left**: Generation performance with generation data controlled at 120k. **Right**: Understanding performance with understanding data controlled at 120k.

beneficial relationship between visual understanding and generation.

More understanding data leads to better generation. Building upon the findings in Section 5.1, we conducted a controlled experiment to investigate whether more understanding data leads to better generation performance. In this experiment, we fixed the generation data to 120K samples and varied the amount of understanding data from 0K to 120K samples, thereby creating models with different levels of understanding. Figure 6a illustrates the overall scores on GenEval, clearly demonstrating that an increase in understanding data correlates with improved generation performance.

More generation data leads to better understanding. To explore the reverse relationship, we conducted another controlled experiment. Here, we fixed the understanding data at 120K samples and vary the amount of generation data across five levels (0K, 30K, 60K, 90K, and 120K). Joint training was performed with the fixed 120K understanding samples. Figure 6b illustrates the models' F1 scores on POPE, demonstrating that increasing the amount of generation data consistently improves understanding performance. This suggests that our generation data positively impacts the model's ability to perform understanding tasks.

Summary. our experiments confirm that, in UnifiedVisualData, generation and understanding data are mutually beneficial. Generation data enhances the model's multimodal understanding, while understanding data improves its generation capabilities. Additionally, we observe that the performance curves in both experiments have not yet converged. This indicates that, by following our data construction process, further scaling of the dataset could lead to even greater performance gains. Moving forward, we plan to expand the dataset to train a more powerful Unified VLLM.

6.2 Reasoning in Multimodal Generation

After training on UnifiedVisualData, Anole-UnifiedVisual demonstrates its ability to effectively leverage reasoning capabilities in visual generation tasks. As illustrated in Figure 9, the model is prompted to generate "an animal associated with having nine lives." While Janus-Pro-7B and Emu3-Gen were trained on larger and higher-quality datasets and can produce more realistic images, they fail to infer that the target animal was a cat. In contrast, Anole-UnifiedVisual successfully deduces that the correct animal is a cat and generates an accurate image. Additional examples are provided in Appendix C. These results indicate that UnifiedVisualData can be used to train models to learn reasoning in multimodal generation.

6.3 Multimodal Reasoning

In the real world, humans often combine mental imagery with textual reasoning to answer questions, as some knowledge is stored more vividly in the form of images in memory. The UnifiedVisualData dataset enhances models' generation capabilities while stimulating their multimodal reasoning abilities by incorporating such multimodal data. For example, in Figure 1, the model is asked, "Which plant has seeds on the outer surface of its fruit". Models like Anole, Janus-Pro-7B, and Emu3-Gen rely on internal knowledge but give incorrect answers. In contrast, the Anole-UnifiedVisual model is capable of effectively "recalling" the appearances of different fruits and combining them to provide the correct answer. This demonstrates that training on the UnifiedVisualData dataset activates multimodal reasoning capabilities in models, allowing them to reason more like humans.

7 Conclusion

In this paper, we propose a novel dataset construction framework, UnifiedVisual, and introduce UnifiedVisualData, a high-quality dataset designed to enhance the synergy between multimodal understanding and generation. Experimental results show that Anole-UnifiedVisual, trained on UnifiedVisualData, consistently outperforms models trained on existing datasets and demonstrates significant mutual enhancement between understanding and generation, fully validating the effectiveness of the UnifiedVisual framework.

Limitations

In this paper, we propose a novel dataset construction framework, UnifiedVisual, and introduce a high-quality dataset, UnifiedVisualData. Through comprehensive experiments, we demonstrate the effectiveness of the dataset. While the current dataset is sufficient to support the experiments and conclusions presented in this paper, it remains relatively small compared to the training datasets used by other open-source models. As demonstrated in Section 6.1, increasing the amount of training data can further enhance model performance. In the future, we plan to leverage the UnifiedVisual framework to construct larger-scale datasets, aiming to further unlock the potential of Unified VLLM.

References

Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. 2023. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2024a. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pages 370–387. Springer.

Wei Chen, Lin Li, Yongqi Yang, Bin Wen, Fan Yang, Tingting Gao, Yu Wu, and Long Chen. 2024b. Comm: A coherent interleaved image-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2406.10462*.

Ethan Chern, Jiadi Su, Yan Ma, and Pengfei Liu. 2024. Anole: An open, autoregressive, native large multimodal models for interleaved image-text generation. *arXiv preprint arXiv:2407.06135*.

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*.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2024. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Preprint*, arXiv:2306.13394.

Tsu-Jui Fu, Wenze Hu, Xianzhi Du, William Yang Wang, Yinfei Yang, and Zhe Gan. 2023. Guiding instruction-based image editing via multimodal large language models. *arXiv preprint arXiv:2309.17102*.

Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. 2024. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint arXiv:2404.14396*.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. 2024. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36.

Seungone Kim, Se June Joo, Doyoung Kim, Joel Jang, Seonghyeon Ye, Jamin Shin, and Minjoon Seo. 2023. The cot collection: Improving zero-shot and few-shot learning of language models via chain-of-thought fine-tuning. *arXiv preprint arXiv:2305.14045*.

Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, et al. 2017. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github. com/openimages*, 2(3):18.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. 2024. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems*, 36.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Hao Li, Changyao Tian, Jie Shao, Xizhou Zhu, Zhaokai Wang, Jinguo Zhu, Wenhan Dou, Xiaogang Wang, Hongsheng Li, Lewei Lu, et al. 2024b. Synergen-vl: Towards synergistic image understanding and generation with vision experts and token folding. *arXiv preprint arXiv:2412.09604*.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023a. Alpacaeval: An automatic evaluator of instruction-following models. `https://github.com/tatsu-lab/alpaca_eval`.

Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. 2024c. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. 2024a. World model on million-length video and language with ringattention. *arXiv e-prints*, pages arXiv–2402.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.

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*, 35:2507–2521.

Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Liang Zhao, et al. 2024. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. *arXiv preprint arXiv:2411.07975*.

Liao Qu, Huichao Zhang, Yiheng Liu, Xu Wang, Yi Jiang, Yiming Gao, Hu Ye, Daniel K Du, Zehuan Yuan, and Xinglong Wu. 2024. Tokenflow: Unified image tokenizer for multimodal understanding and generation. *arXiv preprint arXiv:2412.03069*.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. 2024a. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Yunfan Shao, Linyang Li, Yichuan Ma, Peiji Li, Demin Song, Qinyuan Cheng, Shimin Li, Xiaonan Li, Pengyu Wang, Qipeng Guo, et al. 2024b. Case2code: Learning inductive reasoning with synthetic data. *arXiv preprint arXiv:2407.12504*.

Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2023. Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222*.

Chameleon Team. 2024. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*.

Shengbang Tong, David Fan, Jiachen Zhu, Yunyang Xiong, Xinlei Chen, Koustuv Sinha, Michael Rabbat, Yann LeCun, Saining Xie, and Zhuang Liu. 2024a. Metamorph: Multimodal understanding and generation via instruction tuning. *arXiv preprint arXiv:2412.14164*.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024b. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9568–9578.

Chunwei Wang, Guansong Lu, Junwei Yang, Runhui Huang, Jianhua Han, Lu Hou, Wei Zhang, and Hang Xu. 2024a. Illume: Illuminating your llms to see, draw, and self-enhance. *arXiv preprint arXiv:2412.06673*.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. 2024b. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. 2024. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *arXiv preprint arXiv:2410.13848*.

Penghao Wu and Saining Xie. 2023. V*: Guided visual search as a core mechanism in multimodal llms. *Preprint*, arXiv:2312.14135.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2023. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*.

XAI. 2024. Realworldqa.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. 2024. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*.

Guowei Xu, Peng Jin, Li Hao, Yibing Song, Lichao Sun, and Li Yuan. 2024. Llava-o1: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*.

Lili Yu, Bowen Shi, Ramakanth Pasunuru, Benjamin Muller, Olga Golovneva, Tianlu Wang, Arun Babu, Binh Tang, Brian Karrer, Shelly Sheynin, et al. 2023. Scaling autoregressive multi-modal models: Pretraining and instruction tuning. *arXiv preprint arXiv:2309.02591*, 2(3).

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. 2023. Magicbrush: A manually annotated dataset for instruction-guided image editing. In *Advances in Neural Information Processing Systems*.

Mozhi Zhang, Pengyu Wang, Chenkun Tan, Mianqiu Huang, Dong Zhang, Yaqian Zhou, and Xipeng Qiu. 2024. Metaalign: Align large language models with diverse preferences during inference time. *arXiv preprint arXiv:2410.14184*.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. 2024. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*.

Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. 2024. Multimodal c4: An open, billion-scale corpus of images interleaved with text. *Advances in Neural Information Processing Systems*, 36.

A Prompt template

A.1 Prompt Template for Image Generation

You should generate {number} pairs of instruction and thought about {topic}. Each pair consists of:

Instruction: This instruction requires generating an image. The instruction must only describe the target indirectly, without stating it explicitly (e.g., instead of "Generate an image of a panda," say, "Generate an image of the animal known for its black-and-white fur and its love for bamboo.").

Thought: A detailed reasoning process that interprets the description in the instruction and deduces what should be generated. The thought should make the reasoning explicit and connect the clues to the final answer.

Examples:
{selected examples about this topic}
Do not include the examples in your output.

Table 3: Prompt template used to generate questions and rationales in Topic- and Scene-Based Generation.

Based on the given text, first summarize what image needs to be generated and then convert it into a format suitable for input into DALL·E 3. Just return the input for DALL·E 3, don't return anything else.

Text:
{thought}

Input for DALL·E 3:

Table 4: Prompt template used to generate the DALL-E-3 input in Topic- and Scene-Based Generation.

You will be given an object name. Your task is to:

1. Create an image generation question that:
- Does not directly mention the object name
- Uses related facts, locations, or cultural references to describe it
- Requests the generation of an image
2. Provide a rationale that:
- Explains the logical connection between the facts and the object
- Ends by stating what image should be produced

Output format:
{"question": "[image generation question]","rationale": "[reasoning process and conclusion about the image to generate]"}

Examples:
object name: "the flag of the United States"
{"question": "Show me the national flag of the country where Yellowstone National Park is located.","rationale": "Yellowstone National Park is located in the United States, so the national flag is the American flag. This means we need to create an image of the flag of the United States." }
object name: "the Eiffel Tower"
{"question": "I'd like to see an illustration of the most famous landmark in France, built as the entrance arch for the 1889 World's Fair.","rationale": "The description points to the Eiffel Tower, which was constructed for the 1889 World's Fair and stands as France's most iconic monument. The requested image should be of the Eiffel Tower." }
object name: "a panda"
{"question": "Generates an image of a black and white bear species native to the bamboo forests of central China.","rationale": "The description refers to the giant panda, which is native to China and known for eating bamboo as its main food source. The image we want is of a panda."}

Input:
object name: {Input object name}
Output:

Table 5: Prompt template used to generate questions and rationales in Category- and Image-Based Generation.

A.2 Prompt Template for Image Editing

I will provide you with:
An original image
An instruction for editing the image
An edited image

Your task is:
Based on the given before-edit image, after-edit image, and the editing instructions, analyze the differences between the two images, summarize the most notable features of the after-edit image compared to the original, and describe them in one clear and precise sentence.
It is worth noting that the Main changes include additions, deletions, and modifications, which cannot be expressed explicitly in the Output, but should be expressed implicitly.

Example:
Input:
Original image: A person sitting on a couch in a living room, looking at their phone
Editing instruction: Darken the scene, only keeping the light emitted from the phone screen
Edited image: A person sitting on a couch in a dark room, looking at their phone screen with bright light
Output:
It highlights the light source from the phone screen, creating a dim and focused atmosphere throughout the scene.

Input:
Original image: {The image before editing}
Editing instruction: {The original editing instruction}
Edited image: {The image after editing}
Output:

Table 6: Prompt template used to generate a new editing instruction in Image Editing.

You are a specialized assistant for designing Image editing tasks. I will provide you with:

An original image
Main changes in the image after editing
An edited image
Your task is:
Convert Main changes to a question with answer about the original image that:
1. Can be a request to modify the image or a desired image
2. Must be answered with help from the edited image
3. Must be very relevant to the image and cannot be a general question that has nothing to do with the image
4. It is worth noting that the Main changes include additions, deletions, and modifications, which cannot be expressed explicitly in the question, but should be expressed implicitly.

The answer should use <image_placeholder> to replace the edited image position in the response

Example:
Original image: A person sitting on a couch in a living room, looking at their phone
Edited image: A person sitting on a couch in a dark room, looking at their phone screen with bright light
Main changes: It highlights the light source from the phone screen, creating a dim and focused atmosphere throughout the scene.

Question: How to highlight the light source effect of the mobile phone screen?

Answer:
To highlight the light source effect of the mobile phone screen, we can darken the entire scene while preserving only the light from the phone screen. This will help create contrast and emphasize the phone's light.<image_placeholder>

Input:
Original image: {The image before editing}
Edited image: {The image after editing}
Main changes: {Main changes in the image after editing}
Output:

Table 7: Prompt template used to generate a rationale in Image Editing.

A.3 Prompt Template for Image Correction

I will give you a prompt for image generation. Please help me modify this prompt by changing or removing some key descriptive elements. The modified prompt should create an image that differs from the original in certain visual elements while maintaining the overall theme. Prompt:{generation prompt} Modified prompt:

Table 8: Prompt template used to generate a modified description in Image Correction.

You are a professional image analysis expert. I will provide an image generation requirement and an image generated based on that requirement. This image has some inconsistencies with the original requirements. Please analyze according to these steps: First, carefully analyze the differences and inconsistencies between the image and the requirements. Then, explain in detail how to make adjustments to obtain an image that fully meets the original requirements. End with a phrase similar to "Now, let's generate a new image that fully complies with the requirements based on the above suggestions." Image generation requirement:{generation prompt} Your response:

Table 9: Prompt template used to generate a rationale in Image Correction.

A.4 Prompt Template for MM Reasoning (MM)

[image]
Based on this image, generate a challenging analytical question that has a definitive answer. The question should:

1. Require both careful observation of the image AND application of basic world knowledge
2. Require careful observation and logical reasoning to solve
3. Have a single correct answer rather than subjective interpretations
4. Be specific and precise, not vague or open-ended
5. Use world knowledge that is:
- Commonly understood and easily visualizable
- Not specialized or technical

Just provide the question without any explanation or additional information.
Question:

Table 10: Prompt template used to generate a question based on an image in MM Reasoning (MM).

[image]
You will be given an image and a question. You should analyze the image and answer the question step by step.
The rationale must be in the form of interleaved image descriptions and text. The maximum number of image descriptions in the rationale is 2.
The image descriptions and text in the rationale must complement each other to form a coherent and rigorous chain of reasoning that leads to the correct answer to the question.
The image descriptions in the response are of the form [image: description].
The image descriptions should be simple and concise enough.
The generated image descriptions cannot be close to the original image.
Just return the rationale, don’t return anything else.
Question:{question}
Rationale:

Table 11: Prompt template used to generate a rationale in MM Reasoning (MM).

A.5 Prompt Template for MM Reasoning (T)

Please provide me with a list of {number} questions, options and answers about {topic} for Multiple Choice tasks. These questions must meet the following requirements:
Note that: The questions should have a definite answer. The answer does not change over time. Only one of the options is the correct answer. The questions and answers should not be too related to numbers.
Note that: The questions should be challenging, requiring multiple steps to answer. And the questions should be related to visual information.
Note that: The questions require a chain of thought to deduce the correct answer. The reasoning chain must be in a mixed format of text and descriptions of the images, where the descriptions of the images and text work together to form a coherent and logical chain of reasoning.
{"question": A question generated by you, "options": 4 options in list format generated by you, "answer": The answer generated by you}

Examples:
{selected examples about this topic}

Do not include the examples in your output.
Just provide the questions, options and answers in a jsonline format, without any explanation or additional information.

Table 12: Prompt template used to generate a question in MM Reasoning (T).

You will be given a multiple choice question and its correct answer. You should analyze and answer the question step by step. You need to give the rationale first and finally give the correct answer.
The rationale must be in the form of interleaved image descriptions and text.
The image descriptions and text in the rationale must complement each other to form a coherent and rigorous chain of reasoning that leads to the correct answer to the question.
The image descriptions in the response are of the form [image: description].
Note that: The number of image descriptions in the rationale must be no more than 3.
Note that: The image descriptions of an image should contain the content of only one option.
Note that: The image descriptions should be concise and clear.
Note that: The image descriptions should be easily conveyed visually.
Just return the rationale, don't return anything else.

Question: {question}
Options: {options}
Correct answer: {answer}
Rationale:

Table 13: Prompt template used to generate a rationale based on a question in MM Reasoning (T).

A.6 Prompt Template for Internet Multimodal Data

Given the following interleaved text-image content, please generate a question for which the provided content can serve as the answer.
The images in the provided content are in the form of <image_placeholder>. The question you generated should closely align with the logic of the provided content.
Content: {interleaved content}
Question:

Table 14: Prompt template used to generate a question based on interleaved content in Internet Multimodal Data.

A.7 Prompt Template for Evaluation

You will be provided with a question, its correct answer, and an answer to evaluate. Your task is to determine whether the given answer is correct or not.

Question:
{question}

Correct Answer:
{golden answer}

Answer to Evaluate:
{model output}

Now, determine if the answer to evaluate is correct or wrong and respond only with "Correct" or "Wrong".

Table 15: Prompt template used to determine whether the model's output matches the ground truth.

Here is an answer to a question. This answer may be lengthy, but its final meaning is either "yes" or "no." Please carefully read and summarize the core meaning of this answer, and then determine whether its final answer is "yes" or "no." If the answer does not clearly express "yes" or "no," return "other." You must return only one word: "yes," "no," or "other.
Question:
{question}

Answer:
{model output}

Table 16: Prompt template used to summarize the model's output as 'yes' or 'no'.

B Dataset Construction

B.1 Tools

Tool	Link
GPT-4	https://openai.com/index/gpt-4
GPT-4o	https://openai.com/index/gpt-4o-system-card
DALL-E-3	https://openai.com/index/dall-e-3
text-embedding-ada-002	https://openai.com/index/new-and-improved-embedding-model
clip-vit-large-patch14	https://huggingface.co/openai/clip-vit-large-patch14
stable-diffusion-3.5-large	https://huggingface.co/stabilityai/stable-diffusion-3.5-large
Bing Image Search	https://github.com/hellock/icrawler
Google Custom Search	https://console.cloud.google.com

Table 17: Links to the tools used for constructing UnifiedVisualData.

C Additional Qualitative Results

Figure 7: Examples of multimodal reasoning using Anole-UnifiedVisual.

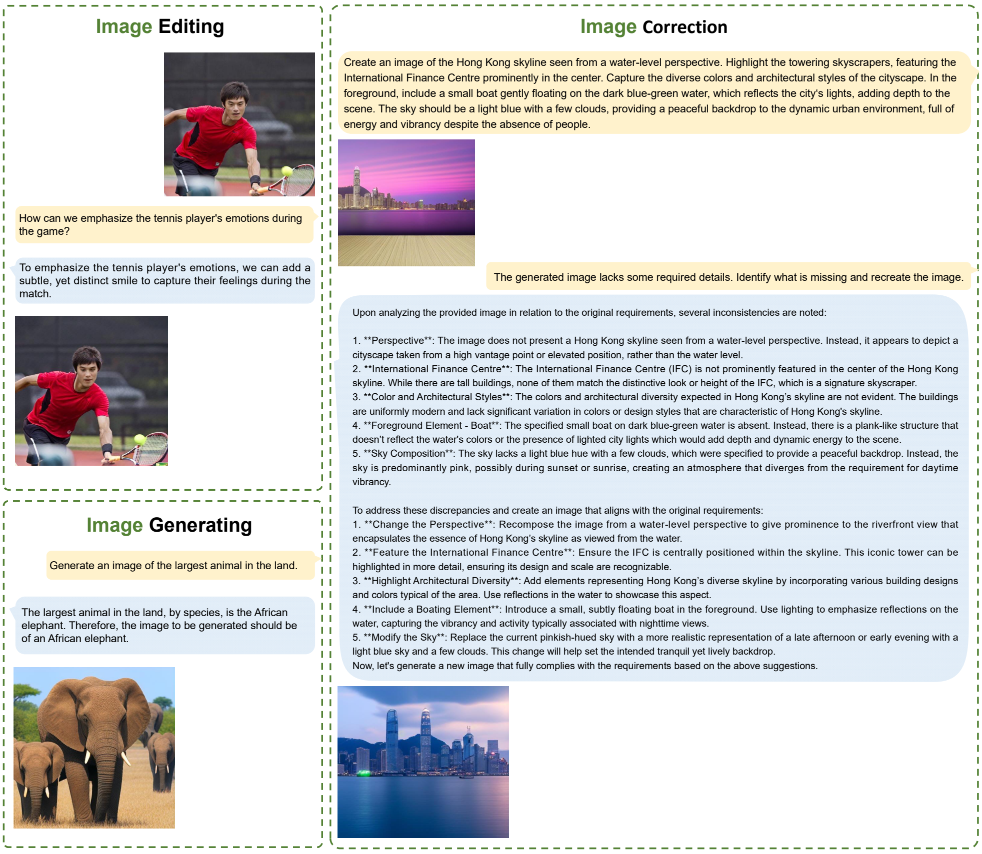


Figure 8: Examples of visual generation using Anole-UnifiedVisual.

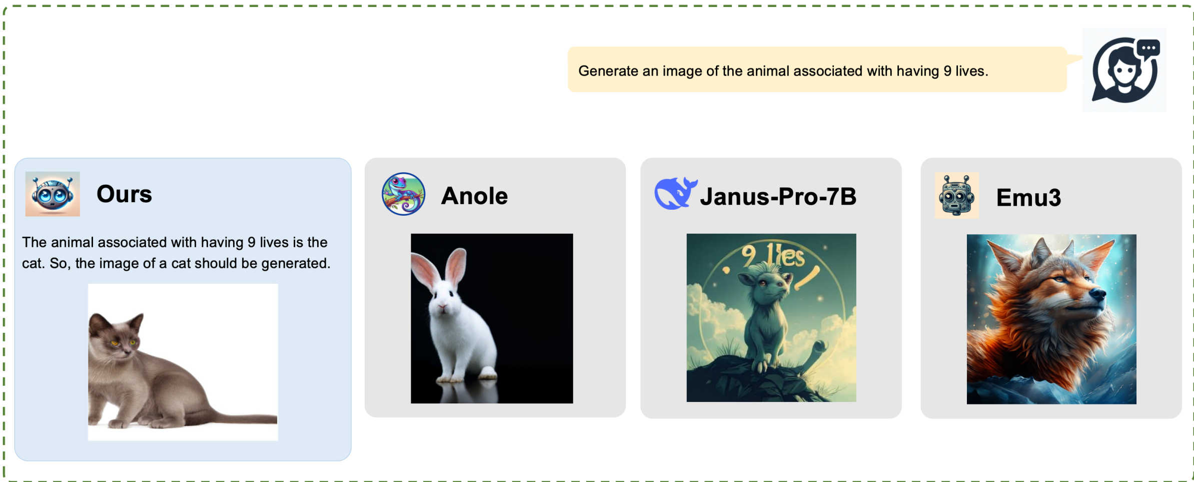


Figure 9: Examples of visual generation using Anole-UnifiedVisual.