

Generative Multimodal Models are In-Context Learners

Supplementary Material

A. Broader Impact and Limitations

Large multimodal models offer a wide range of benefits to society, from enhancing visual navigation and medical diagnostics to increasing accessibility for individuals with visual impairment. The in-context learning capabilities of **Emu2** allow it to quickly adapt to new tasks or environments, even with limited data, ushering in numerous potential applications. The generative capabilities of **Emu2** can be highly valuable to the creative industries.

However, there are potential downsides in more powerful multimodal models to be considered. The hallucination issue of multimodal models may cause incorrect and unreasonable predictions in certain cases. **Emu2** may also generate harmful or biased content like other generative models [55] since the training data may be biased or contain unsuitable content. We are actively working to enhance the robustness of multimodal models, reduce model hallucinations, improve the fairness of training data, and reduce toxic data. We also call on the wider community to pay attention to the potential social impact of multimodal models as they are growing larger and stronger.

One of the limitations of **Emu2** is that its in-context learning capability could fail in some complex scenes or tasks, *e.g.*, counting in a crowd. Additionally, there is still a gap between **Emu2**'s question-answering capability and that of closed multimodal systems. For example, GPT-4V achieves 67.7 MM-Vet score vs. **Emu2**'s 48.5, although already being state-of-the-art among public models. We believe there is much room to improve as the quality and quantity of training data improve and as model scale continues to grow.

B. More Pretraining Details

B.1. Dataset Details

In pretraining, we exclusively leverage image-text pairs and video-text pairs for stage 1 training. We additionally leverage interleaved and language-only data altogether for stage 2. The integration of visual embeddings with text tokens generates unified multimodal sequences. These sequences are then structured by appending the tokens $\langle s \rangle$ and $\langle /s \rangle$ to denote the beginning and end of each sequence.

Image/Video-text Pairs. In the pretraining stage, we utilize image-text pairs from LAION-2B [65] and CapsFusion-120M [87], along with video-text pairs from WebVid-10M [8]. During pretraining stage 2, each image or video is randomly placed before or after its corresponding text

with a probability of 0.5, respectively. For each video, we randomly sample 8 frames. To structure the visual embeddings, we append two special tokens, $[\text{IMG}]$ and $[/ \text{IMG}]$, to signify the start and end of the visual embeddings. In the case of videos, where there are T frames, each frame is encoded into a set of visual embeddings, and a special token, $[\text{VIDEO}]$, is prepended to the start of the frame embedding sequence. This design helps distinguish between multiple images and video frames within the multimodal sequences.

Interleaved Image/Video-text Data. We harness the Multimodal-C4 (MMC4) dataset [95] and the YT-Storyboard-1B dataset [71] as expansive sources of image and video-text interleaved data. This approach aims to unlock the in-context learning capability of multimodal models. For each MMC4 document, we randomly sample $N = 8$ images, accompanied by their corresponding sentences, to construct a subsequence of $L = 1024$. During pretraining stage 2, each image or frame is randomly positioned before or after its corresponding text with a probability of 0.5. The special tokens used in this interleaved data are consistent with those employed in the image-text pair data.

Grounded Image-text Pairs. We curated a dataset of grounded image-text pairs named CapsFusion-grounded-100M, employing data from CapsFusion [87] processed through the dataset construction pipeline proposed by Kosmos-2 [57]. Additionally, we utilized the 20M GRIT dataset introduced by Kosmos-2 [57]. To enhance the diversity and context of the dataset, we randomly positioned each phrase before or after its corresponding coordinates with a probability of 0.7. The bounding box can be represented using its top-left point (x_1, y_1) and bottom-right point (x_2, y_2) . We transform continuous coordinates into 224 discrete tokens [57], the coordinates of a sample box can be formulated as $\langle loc_{000} \rangle \langle loc_{000} \rangle \langle loc_{224} \rangle \langle loc_{224} \rangle$. We added these tokens to the word vocabulary to facilitate unified modeling with text. To distinguish grounding text from regular text strings, we introduced two special tokens, $\langle \text{COOR} \rangle$ and $\langle / \text{COOR} \rangle$, marking the beginning and end of the bounding box coordinates. Moreover, to establish the correct association between bounding boxes and their corresponding descriptive phrases, an additional set of special tokens, $\langle p \rangle$ and $\langle / p \rangle$, was appended. To guide the model in grounding text output to the provided image, we utilized the special token $\langle \text{grounding} \rangle$. This comprehensive set of tokens and instructions enriches the training data for effective multimodal modeling and understanding.

Language-only Data. To maintain text reasoning capabilities, we engage in joint training with the language modeling dataset Pile [26]. The entire text corpus from Pile is preprocessed offline, and each training sample is tokenized into 2048 tokens using the LLaMA tokenizer. We randomly sample a total of 3.6 billion tokens for pretraining purposes.

B.2. Training Hyperparameters

We report the detailed training hyperparameter settings of **Emu2** during the pretraining in Table 6.

Configuration	Emu2 Stage 1	Emu2 Stage 2
Visual Encoder init.	EVA-02-CLIP-E-plus	Emu2 stage 1
Multimodal Modeling init.	LLaMA-33B	Emu2 stage 1
Linear projection layer init.	random	Emu2 stage 1
Input image resolution	224 ² 448 ²	448 ²
Optimizer	AdamW	
Optimizer hyper-parameters	$\beta_1 = 0.9, \beta_2 = 0.95, eps = 10^{-6}$	
Peak learning rate	$1 \times 10^{-4}, 3 \times 10^{-5}, 5 \times 10^{-5}$	1×10^{-5}
Learning rate schedule	cosine decay	
Gradient clip	5.0	
Training steps	35.2k 4.0k	20.35k
Warmup ratio	0.02	0.1
Global batch size*	6144, 768	12800, 6400, 3200, 800
Numerical precision	bfloat16	

Table 6. Summary of pretraining hyperparameters of **Emu2** in pretraining stages. Peaking learning rates are 1×10^{-4} for the linear projection layer, 3×10^{-5} for Multimodal Modeling, and 5×10^{-5} for Visual Encoder. *Global batch size: 1) 6144 for image-text pairs and 768 for video-text pairs in stage 1. 2) 12800 for image-text pairs, 6400 for video-text pairs, 3200 for image-text/video-text interleaved data, and 800 for language-only data in stage 2.

B.3. Visual Decoding

B.3.1 Dataset Details

We utilize images in LAION-COCO [2] and LAION-Aesthetics [1] to train the Visual Decoder. Images whose resolution is smaller than 512×512 are filtered to prevent generating low-quality results. We employ ratio-preserving random scaling followed by random cropping of a square portion from the scaled image to keep all training images unstretched. The original image size and crop coordinates are used as additional conditions following SDXL [58].

B.3.2 Training Hyperparameters

The detailed hyperparameters of visual decoding training are summarized in Table 7.

Configuration	Visual Decoding
Visual Encoder init.	Emu2 stage 1
Visual Decoder init.	SDXL-base
Encoder input image resolution	448 \times 448
Decoder output image resolution	1024 \times 1024
Optimizer	AdamW
Optimizer hyper-parameters	$\beta_1 = 0.9, \beta_2 = 0.999, eps = 10^{-8}$
Peak learning rate	1×10^{-4}
Learning rate schedule	log warm-up, linear decay
Gradient clip	1.0
Total training steps	8,000
Warmup steps	2,500
batch size	2,048
Numerical precision	bfloat16
Classifier-free guidance	10%
Noise offset	0.1

Table 7. Summary of training hyperparameters of **Emu2** Visual Decoder. The Visual Encoder is frozen during training.

C. Instruction-Following Chat

C.1. Dataset Details

We used two types of training data, academic-task-oriented data and multi-modal chat data, in instruction fine-tuning of **Emu2-Chat**. The academic-task-oriented datasets we utilized comprise image captioning datasets such as COCO Caption [18], and TextCaps [66], as well as visual question-answering datasets like VQAv2 [28], OKVQA [53], GQA [32], TextVQA [68], and multi-modal classification data constructed in M3IT [45]. RefCOCO [34], RefCOCO+ [52] and RefCOCOg [52] datasets are also used. The public multi-modal chat data we use includes GPT-assisted visual instruction data LLaVa [49] and LLaVaR [93], language instruction data from ShareGPT [4] and Alpaca [72], and video instruction data from VideoChat [44]. Beyond these, we constructed instruction fine-tuning data from an analog clock reading dataset [83]. For academic-task-oriented datasets, we use the system message “You are a helpful assistant, dedicated to provide concise and efficient answers.”, and for the multi-modal chat data, the system message is “You are a helpful assistant, dedicated to delivering comprehensive and meticulous responses.”.

C.2. Training Hyperparameters

The detailed training hyper-parameters of **Emu2-Chat** are summarized in Table 8.

D. Controllable Visual Generation

D.1. Dataset Details

We use the grounded image-text pairs dataset, *i.e.*, CapsFusion-grounded-100M and GRIT [57] for

Configuration	Emu2-Chat
init.	Emu2
Input image resolution	448 × 448
Optimizer	AdamW
Optimizer hyper-parameters	$\beta_1 = 0.9, \beta_2 = 0.98, eps = 10^{-6}$
Peak learning rate	1×10^{-5}
Learning rate schedule	cosine decay
Gradient clip	5.0
Training steps	8,000
Warmup steps	100
Global Batch size	768
Numerical precision	bfloat16

Table 8. Summary of training hyperparameters of **Emu2-Chat**.

grounded text-to-image generation. We use SAM [36] to obtain segmentation results for the corresponding grounding boxes. We leverage InstructPix2Pix constructed by [10] for image editing tasks. The sample will be formulated as “<s>[IMG]embedding of origin image[/IMG]instruct editing prompt[/IMG]embedding of edited image[/IMG]</s>”. For the text-to-image task, we use a filtered subset the CapsFusion [87], LAION-Aesthetics [1], SA-1B [36], and LAION-High-Resolution [3].

For high-quality fine-tuning, our datasets were meticulously sourced from premium sources, *e.g.*, Unsplash [20], and outputs from advanced text-to-image systems, *e.g.*, Midjourney-V5 [54] and DALL-E-3 [9]. This comprehensive approach ensured a diverse and rich dataset, comprising approximately 500,000 instances of high-quality image-text pairs, instrumental in refining and enhancing the aesthetic quality of our **Emu2-Gen** model’s generated images.

D.2. Training Hyperparameters

We report the detailed training hyperparameter settings of **Emu2-Gen** during the instruction-tuning in Table 9.

E. Evaluation Details

Pretrained Base Model. For few-shot evaluation of **Emu2**, we adopt the Retrieval In-Context Example Selection (RICES) approach for choosing few-shot examples, following Flamingo [5] and Emu [71]. The chosen few-shot examples will be separated by “. ” and then placed ahead of the test sample. We use the prompt “[image] based on the picture, [question] short answer:”.

For zero-shot evaluation, as no example is given, we find the above simple prompt cannot effectively control the model behavior and the model tends to output a sentence rather than a word or phrase. Thus, we modify the prompt to “[image] based on the picture, answer in one word or phrase. [question] short answer:”. This adjustment aligns the model’s output more closely with the distribution of the tested datasets, where responses

Configuration	Emu2-Gen stage1	Emu2-Gen QFT
init.	Emu2	Emu2-Gen stage1
Input image resolution	448 × 448	
Optimizer	AdamW	
Optimizer hyper-parameters	$\beta_1 = 0.9, \beta_2 = 0.95, eps = 10^{-6}$	
Peak learning rate	5×10^{-5}	1×10^{-5}
Learning rate schedule	cosine decay	
Gradient clip	1.0	
Training steps	3k	0.9k
Warmup ratio	0.0	
Global Batch size*	4096, 3584, 2048	2048, 1024, 2048
Numerical precision	bfloat16	

Table 9. Summary of training hyperparameters of **Emu2-Gen**.

*Dataset types are text-to-image pairs, grounded text-to-image and image editing pairs.

typically consist of a succinct word or phrase. The splits and metrics for each benchmark are detailed in Table 10.

Instruction-Following Chat. The evaluation of **Emu2-Chat** follows the assessment method of Emu-I [71], utilizing generation hyper-parameters with a beam size of 5. For video input, 16 frames are uniformly sampled as visual conditions. In the question-answering benchmark that requires short answers, we employ the system message “You are a helpful assistant, dedicated to provide concise and efficient answers.” along with the output format control information used in [48]. In the benchmark for scoring with GPT-4, we use the system message “You are a helpful assistant, dedicated to delivering comprehensive and meticulous responses.”. We provide an overview of the evaluation benchmarks in Table 10.

Benchmark	Task	Split	Metric
VQAv2	Scene understanding VQA	Test-dev	VQA score(↑)
VizWiz	Scene understanding VQA	Test-dev	VQA score(↑)
GQA	Understanding & reasoning VQA	Test-dev	EM(↑)
OKVQA	External knowledge VQA	Val	VQA score(↑)
TextVQA	Text-oriented VQA	Val	VQA score(↑)
Hateful Memes	Meme classification	Seen Test	ROC AUC(↑)
RefCOCO	Refer expression comprehension	-	Accuracy(↑)
RefCOCO+	Refer expression comprehension	-	Accuracy(↑)
RefCOCog	Refer expression comprehension	-	Accuracy(↑)
MSVD-QA	Event understanding VQA	Test	EM(↑)
MSRVTT-QA	Event understanding VQA	Test	EM(↑)
MMMU	Massive multi-discipline QA	Test	Accuracy(↑)
SEED-Bench	Image/Video multi-choice QA	-	Accuracy(↑)
MM-Vet	Open-ended generation	-	GPT-4 score(↑)
TouchStone	Open-ended generation	-	GPT-4 score(↑)

Table 10. Summary of the evaluation benchmarks.

Controllable Visual Generation. For all evaluation of visual generation tasks, we use EulerDiscreteScheduler [33]

with 50 diffusion steps. The classifier-free guidance scale is set to 3.0. To evaluate on DreamBench [63], we select exactly the same image for each object as chosen in Kosmos-G [56]. Similarly to Kosmos-G, we also slightly modified the original prompt with the prefix "a", for example, "a red {}" is modified to "{} Make it red"

F. Qualitative Results

We present qualitative cases for **Emu2-Gen** in Figure 5-11 and for **Emu2-Chat** in Figure 12-14, respectively.

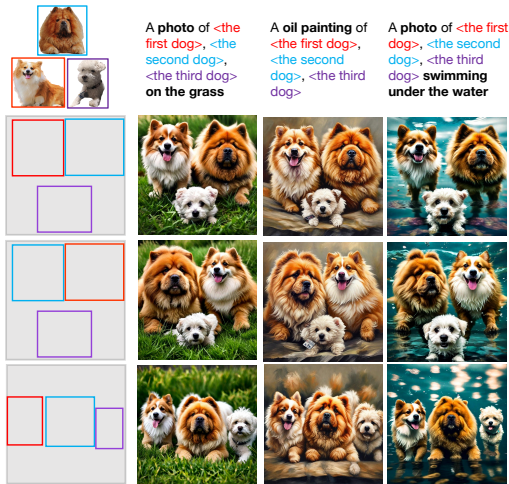


Figure 5. Illustration of controllable visual generation of subject-driven generation across multiple images with layout guidance.

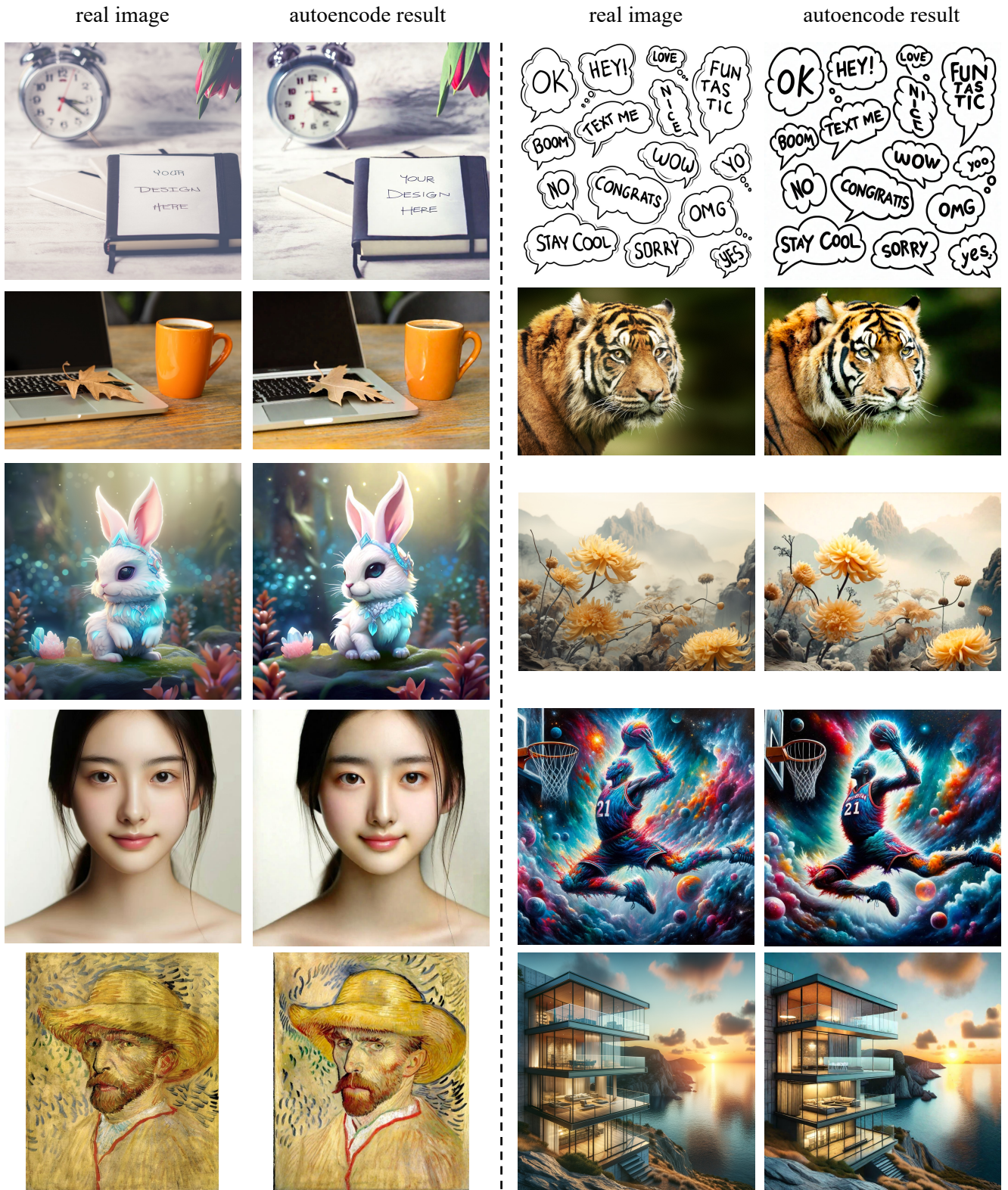


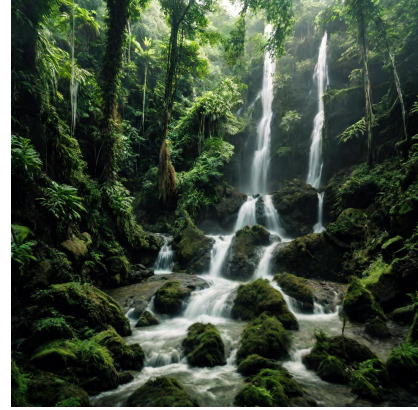
Figure 6. Qualitative cases of image autoencoding.



Hand-painted castle, surrounded by trees, autumn, leaves flying



A dog that has been meditating all the time.



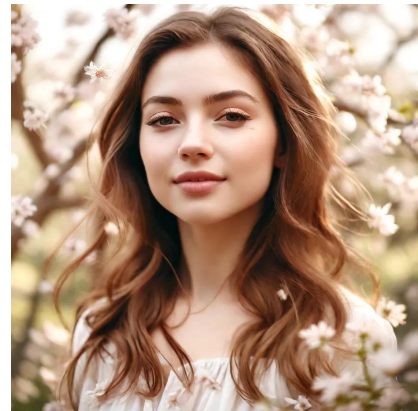
Majestic waterfall in a lush, green tropical forest.



Three minuscule explorers dwarfed by Ethereal monumental gigantic cliffs, minimalistic landscape, beige and gray snowscene , in the style of Andy Fairhurst.



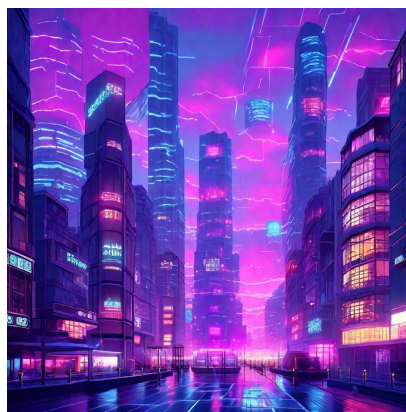
A Blue bird



A woman with natural beauty, her hair gently flowing, and a soft, genuine smile, standing in a real-world cherry blossom park.



In this scene, a cyclist pedals through a sun-dappled forest trail, with rays of light filtering through the dense foliage. The vibrant colors of the rider's gear harmonize with the natural hues of the surrounding flora.



cyber punk city street,all building made from translucent hologram glass material.pastel neon light,night scene with pastel purple and pink sky



A panoramic view of the Grand Canyon at sunset, with the vast, layered rock formations dramatically lit by the fading golden sun, casting deep shadows and highlighting the rich, red hues of the canyon walls.

Figure 7. Illustration of text-to-image generation.

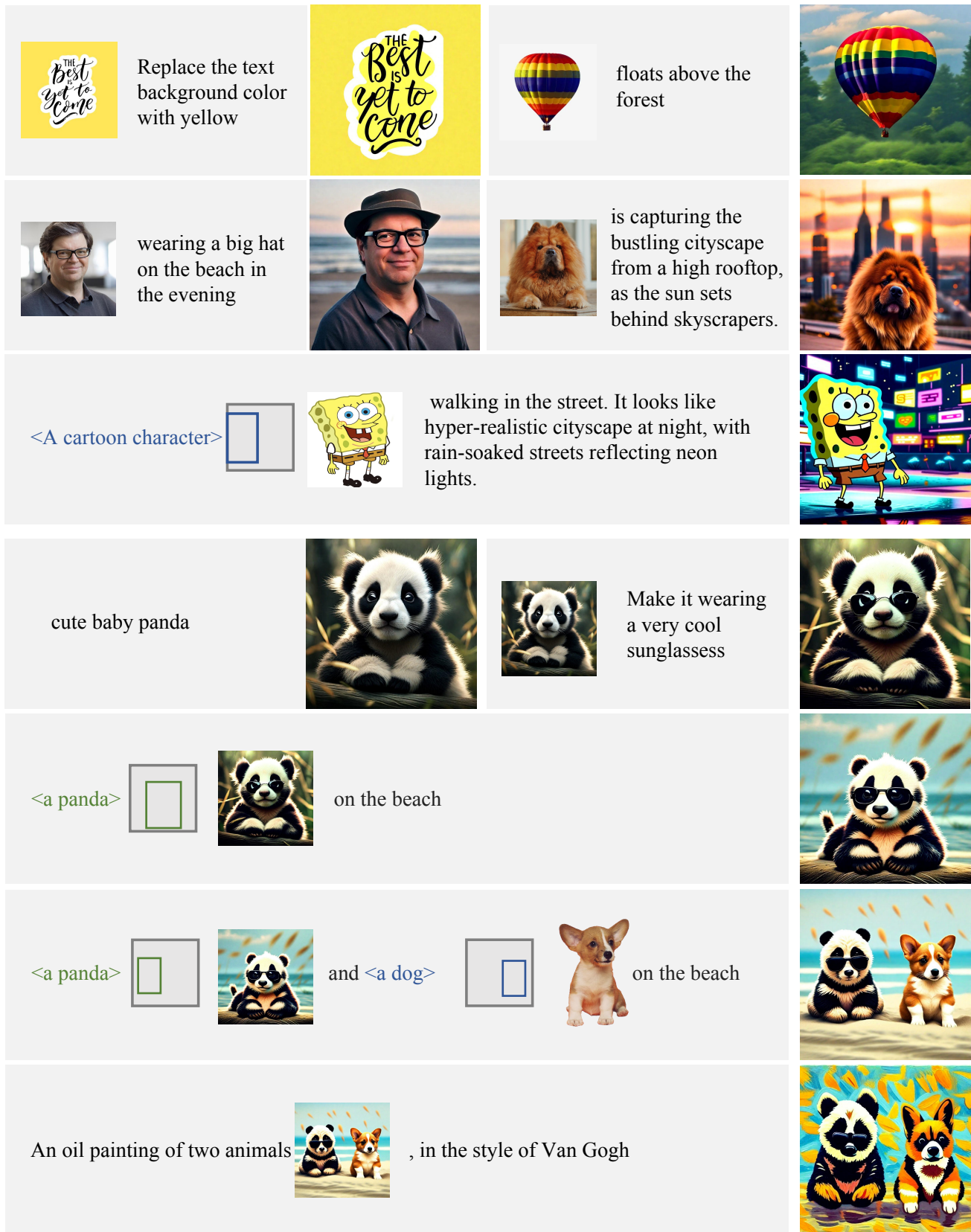


Figure 8. Illustration of zero-shot controllable visual generation with interleaved vision-language prompt.















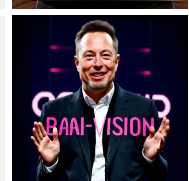




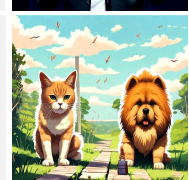



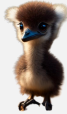
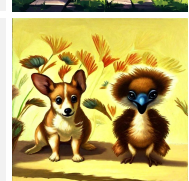






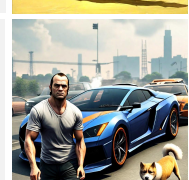












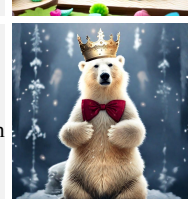
<p><a pekingese>  is on the beach with an <umbrella>  near the <sea> </p>	
<p><A singer>  in <a black cloth>  is singing into <microphone> </p>	
<p><a dog>   in the cover of <a book>   on the table</p>	
<p>A photo of <a man>   holding the <words on neon lamp>  BAAI-VISION on the stage</p>	
<p>An anime draw of <a cat>   and <a dog>   in the style of </p>	
<p>An image draw of <a dog>   and <an emu>   in the style of </p>	
<p><A man>   with <a dog>  , and <a car>   in GTA-V</p>	
<p>A photo of a birthday party in the garden, <a cat>   wearing <a hat>   and eating <a cake>  </p>	
<p><a bear>   wearing <a crown>   and <a bow tie>   in the snow kingdom</p>	

Figure 9. Illustration of zero-shot controllable visual generation with interleaved vision-language prompt.

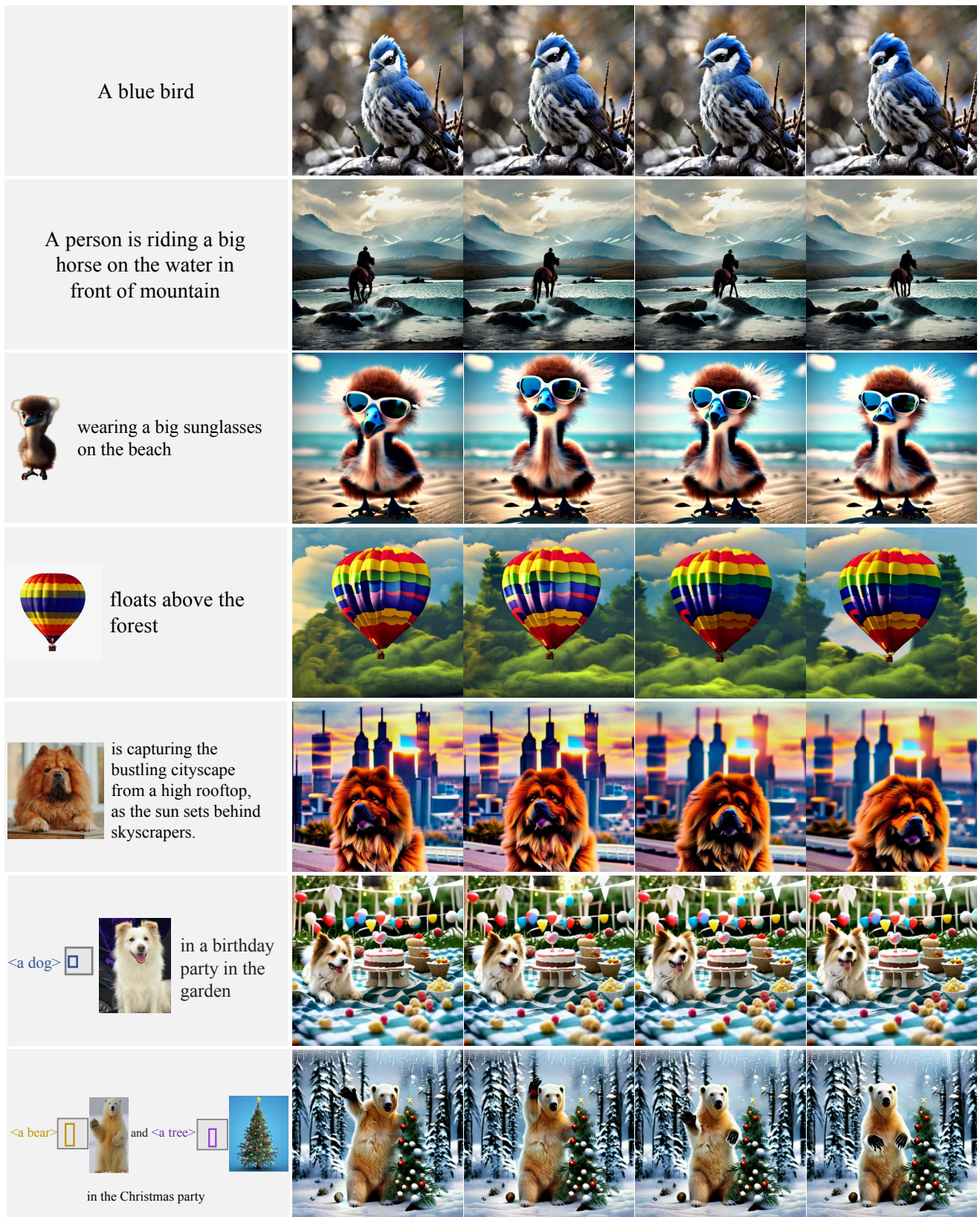


Figure 10. Illustration of zero-shot video generation with interleaved vision-language prompt.

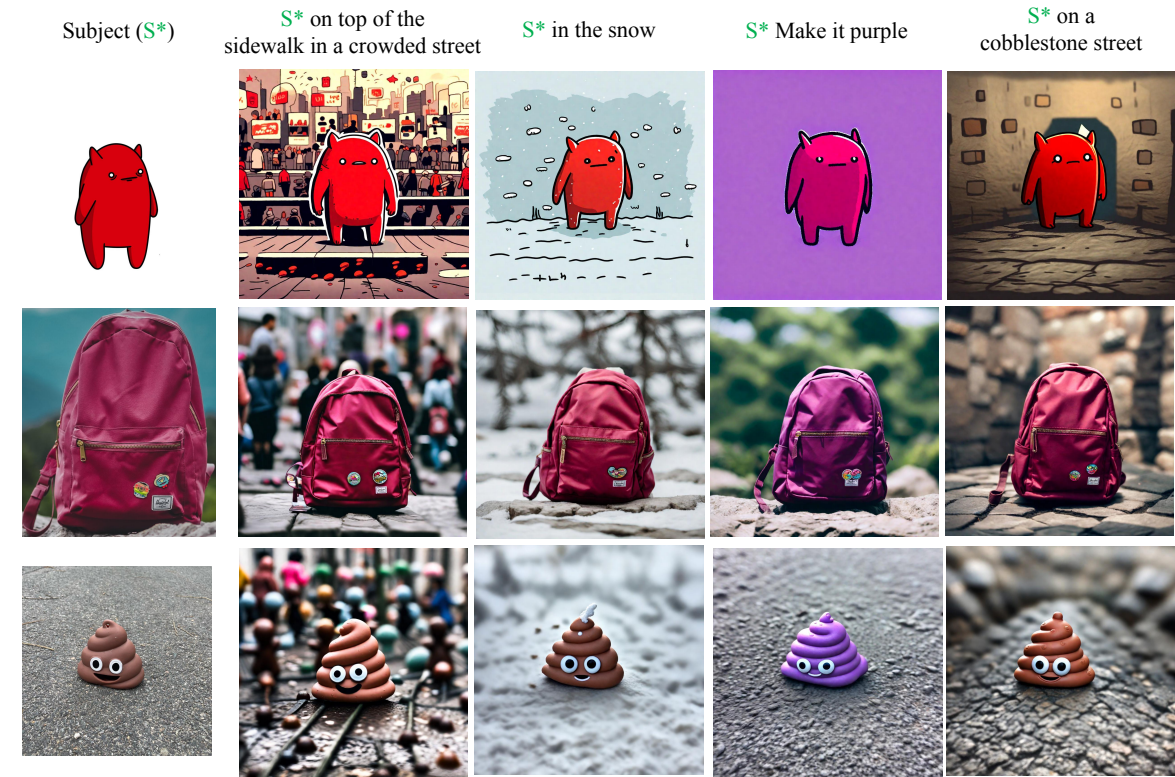
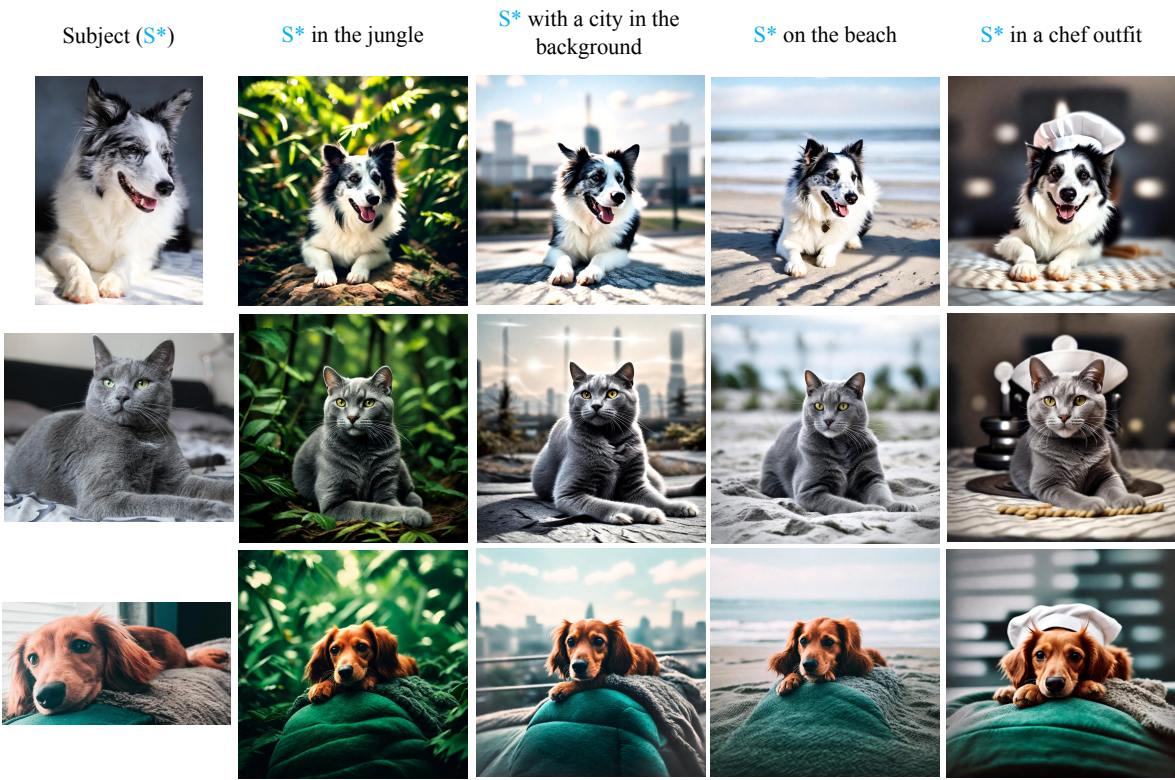
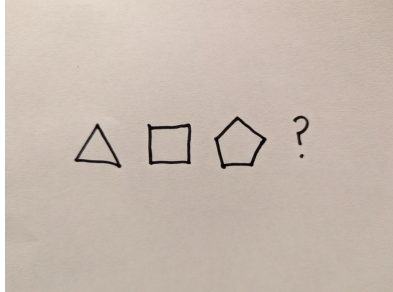


Figure 11. Zero-shot subject-driven generation results on DreamBench.

Input Image



Input Text

Look at this sequence of three shapes. What shape should come as the fourth shape? Explain your reasoning with detailed descriptions of the first shapes.

Model Output

Emu2-Chat

The first shape is a triangle, the second shape is a square, and the third shape is a pentagon. To find the next shape in the sequence, we need to identify a pattern. In this case, the pattern is that each shape has one more side than the previous shape. Therefore, the next shape in the sequence should have six sides, which is a hexagon.

CogVLM-Chat

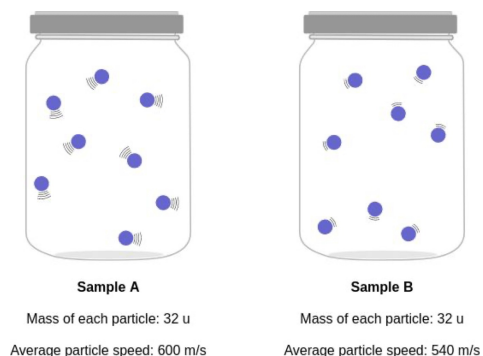
The sequence of shapes could be interpreted in various ways, but one possible continuation is a square, as it complements the triangle and pentagon shapes by forming a 2D shape that can be thought of as a 'flat' version of the other three. The triangle represents a 3D shape, the pentagon a 5D shape, and the square a 2D shape.

Gemini Ultra

The fourth shape should be a hexagon. The first shape is a triangle, the second shape is a square, and the third shape is a pentagon. The number of sides in each shape is increasing by one

Figure 12. Qualitative case of **Emu2-Chat**.

Input Image



Input Text

Compare the average kinetic energies of the particles in each sample. The diagrams below show two pure samples of gas in identical closed, rigid containers. Each colored ball represents one gas particle. Both samples have the same number of particles. Which sample has the higher temperature? Sample A or sample B?

Model Output

Emu2-Chat

The sample with the higher temperature is Sample A, where the mass of each particle is 32 u and the average particle speed is 600 m/s. In Sample B, the mass of each particle is 32 u, but the average particle speed is 540 m/s. Since the temperature is a measure of the average kinetic energy of the particles, and kinetic energy is proportional to the square of the velocity, the higher the velocity, the higher the temperature. In Sample A, the higher average particle speed results in a higher temperature compared to Sample B.

CogVLM-Chat

The kinetic energy of a particle is given by the formula $(1/2)mv^2$, where m is the mass and v is the velocity of the particle. Since the velocities of the particles in both samples are the same, the sample with the higher mass particles will have the higher kinetic energy. Therefore, Sample A, with particles having a mass of 32 u, will have a higher average kinetic energy than Sample B, which has particles with a mass of 32 u.

GPT-4V

Because the average particle speed in sample A is higher than in sample B (600 m/s vs. 540 m/s), the particles in sample A have a higher average kinetic energy. Therefore, sample A has a higher temperature than sample B.

Figure 13. Qualitative case of **Emu2-Chat**.

Input Image



Input Text

Imagine you are a guiding robot.
Here is a photo I took. Please tell me how to get to the restroom.

Model Output

Emu2-Chat

To get to the restrooms, you should follow the "Restrooms" sign in the picture. The sign is pointing to the left, so you should head in that direction to find the restrooms. If you need further assistance, please let me know, and I'll do my best to help you navigate to the restrooms.

CogVLM-Chat

To get to the restroom, follow the direction indicated by the sign. It's located to the right of the sign and is labeled 'Restrooms'.

Qwen-VL-Chat

To reach the restrooms, follow the directional arrow on the sign pointing downwards and towards your right side. The restroom should be located in that direction based on the information provided by the sign. Keep an eye out for any additional signs or markers indicating where the restrooms might be within this area of the building.

Figure 14. Qualitative case of **Emu2-Chat**.