ILLUME: Illuminating Your LLMs to See, Draw, and Self-Enhance Supplementary Materials

Chunwei Wang[1*], Guansong Lu[1*], Junwei Yang[1*], Runhui Huang[2],
Jianhua Han[1], Lu Hou[1], Wei Zhang[1], Hang Xu[1†]
[1]Huawei Noah's Ark Lab, [2]The University of Hong Kong

A. More Implementation Details

Vision Tokenizer. We supervise the quantization process within a semantic feature space, which is promising to facilitate the image-text alignment in MLLM training. Given an image x, it is fed into vision encoder to extract semantic features $X = \{x_1, ..., x_N\}$. The semantic features then pass into a quantizer, which tokenizes X to a sequence of discrete tokens $V = \{v_1, ..., v_N\}$ by looking up a learnable codebook $\mathcal{C} = \{c_1, ...c_K\}$, where K is the codebook size. The discrete token v_i is calculated by assigning x_i to its closest neighbourhood code in $\mathcal{C}$ according to the L2 norm:

$$v_i = \arg\min_j ||x_i - c_j||, v_i \in [0, K-1] \tag{1}$$

Based on the discrete tokens, we can obtain its quantized embeddings, which is then fed into a decoder to obtain reconstructed semantic features $X^{rec} = \{x_1^{rec}, ..., x_N^{rec}\}$. The quantization process is supervised by the feature reconstruction loss using *cosine loss* and *smoothl1 loss*:

$$\mathcal{L} = \sum_{i=1}^{N}(smoothl1(x_i, x_i^{rec}) + (1 - cosine(x_i, x_i^{rec}))) \tag{2}$$

During training, the vision encoder is kept frozen and only the parameters of quantizer and decoder are updated. It is trained for 80K steps on 80M images with the batch size of 2048, epochs of 2 and learning rate of 5e-5.

To further recover the original pixel space, the reconstructed semantic features are set as conditions and injected to each block of a conditional diffusion model through cross-attention layers. The conditional U-Net is initialized from SDXL and finetuned 80K steps with the batch size of 128 and learning rate of 2e-5. Only the attention layer of U-Net is updated for efficient training. Note that the whole tokenizer training only requires pure image data without corresponding text descriptions.

More Explanation of MLLM Framework. In our MLLM framework, we employ a continuous-input discrete-output scheme for visual images. Figure A illustrates an example of data sequence under this scheme, where all the data are formatted in QA pairs and we only compute loss and conduct teacher forcing on the answer. The main reason behind this mechanism is to avoid information loss due to vector quantization and improve the visual understanding capability of unified models.

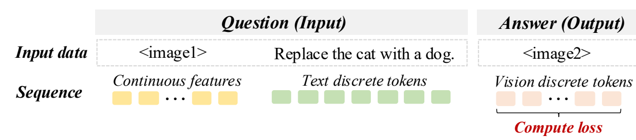


Figure A. An example of data sequence in editing task. The input images are processed through a vision encoder and an adapter while images in the answer are represented as discrete tokens.

Figure B. Comparison of different hyper-parameters in inference.

B. More Results of ILLUME

More Qualitative Results. Figure D showcases additional qualitative results for comprehension tasks, demonstrating that our ILLUME model can adeptly handle various comprehension tasks and images with significant differences in aspect ratio. Figures E and Figures F provide further visualizations in text-to-image generation and mixed-modal generation tasks, respectively. In the future, we plan to enhance MLLMs to produce higher resolution images and to support a wider range of mixed-modal generation tasks.

Method	Params.	Type	*Basic*						*Advanced*					
			Attribute	Scene	Spatial	Action	Part	Overall	Count	Differ	Compare	Negate	Universal	Overall
SDXL	2.6B	Diffusion	0.84	0.84	0.82	0.83	0.89	0.83	0.71	0.73	0.69	0.50	0.66	0.63
LWM	7B	Autoregressive	0.63	0.62	0.65	0.63	0.70	0.63	0.59	0.58	0.54	0.49	0.52	0.53
Show-o	1.5B	Autoregressive	0.72	0.72	0.70	0.70	0.75	0.70	0.70	0.62	0.71	0.51	0.65	0.60
VILA-U(256)	7B	Autoregressive	0.78	0.78	0.77	0.78	0.79	0.76	0.70	0.71	0.74	0.53	0.66	0.64
VILA-U(384)	7B	Autoregressive	0.75	0.76	0.75	0.73	0.75	0.73	0.68	0.67	0.71	0.51	0.64	0.61
ILLUME (Ours)	7B	Autoregressive	0.75	0.79	0.75	0.77	0.73	0.75	0.66	0.68	0.67	0.49	0.63	0.60

Table A. Detailed quantitative results on GenAI-bench.

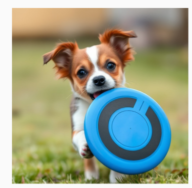
[Turn 1]
User:
Text description: A tiny brown dog with white patches, eagerly holding a blue and black Frisbee in its mouth, ready to play
Image: {Output image}
Please identify the inconsistencies between the image and text, and analyze them from the following aspects**: Object, Counting, Color, Spatial Relation.**

Give the score and corresponding analysis for each aspect. Please answer in JSON format.

Assistant:
{"**Object**": {"Score": 1, "Analysis": "The text describes a dog and a Frisbee, both of which are present in the image. "},
"**Counting**":{"Score": 1,"Analysis": "No relation"},
"**Color**":{"Score": 0,"Analysis": "**The text describes a brown dog with white patches and a blue and black Frisbee, while the image shows a brown dog without white patches and a multicolored Frisbee.**"**},**
"**Spatial Relation**":{"Score": 0,"Analysis": "**The text describes the dog holding the Frisbee in its mouth, but the image shows the Frisbee on the ground.**"}}

[Turn 2][Additionally for bad generated images]
User:
Based on the above analysis, carefully generate an image according to the text description: A tiny brown dog with white patches, eagerly holding a blue and black Frisbee in its mouth, ready to play. Please make sure the content of generated image is consistent with the text.
Assistant:
{GT image}

Output image



GT image

Figure C. Data example of assessment data for self-enhancing multimodal alignment.

Encoder	POPE	MME-P	SEED	AI2D	ChartQA	DocVQA	InfoVQA	OCRBench
SigLIP	87.6	1404.5	65.5	55.1	54.9	60.4	38.7	414
UNIT	87.6	**1452.6**	**65.8**	**55.2**	**56.3**	**61.4**	**40.3**	**440**

Table B. Comparison of UNIT and SigLIP.

Inference Hyper-parameters. Figure B presents a comparison of different inference decoding hyperparameters for text-to-image generation. It can be observed that increasing temperature, top-k, and guidance scale all lead to improved visual details.

Detailed Performance Results on GenAI-bench. We details per-category performance on GenAI-bench in Table A, where our ILLUME achieves competitive results with current autoregressive-based unified MLLMs.

Choices of Vision Encoder. As illustrated in Table B, we adopt UNIT as vision encoder due to its superior performance over SigLIP on document/OCR benchmarks.

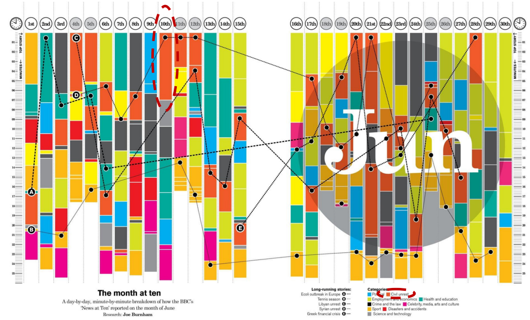
More Ablation of Self-enhancing Multimodal Alignment. We use GPT-4o to generate assessment labels purely to ensure data quality and Table C shows that replacing GPT-4o with Qwen2.5-VL-72B yields consistent performance gains, indicating the robustness of our approach. We believe that as the unified model becomes stronger (currently based on Vicuna-7B) or with proper data filtering process, it can eventually serve as its own assessor.

Source of Assessment	*Understanding*						Generation
	POPE	MME-P	SEED	GQA	MM-Vet	MMMU	Geneval
baseline	86.4	1358.6	61.7	60.0	27.4	31.2	0.56
Qwen2.5-VL-72B	86.0	1429.6	65.9	61.0	30.3	31.9	0.58
GPT-4o	86.1	1446.7	66.0	60.7	29.0	32.0	0.59

Table C. Ablations of different source of the assessment labels.

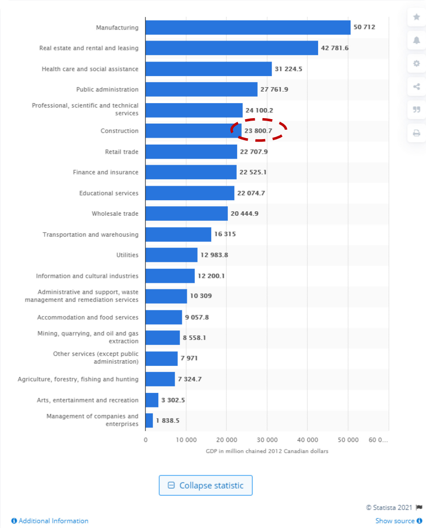
Data Examples of Assessment Data. Figure C illustrates an example of assessment data for self-enhancing multimodal alignment scheme. This example showcases how the data identifies specific reasons for inconsistencies between self-generated images and text descriptions, which aids the model in interpreting images more accurately and helps prevent mistakes during image generation.

Q: What percent of non-digitally mature healthcare providers do not use digital technologies to personalize care as per the survey?
A: 79%



Q: What category was the top story on the 10th of June?
A: civil unrest

Q: How many hot air balloons are in the image?
A. 1
B. 3
C. 2
D. 4
A: B



Q: What was the GDP of the construction industry in Quebec in 2012 Canadian dollars?
A: 23800.7

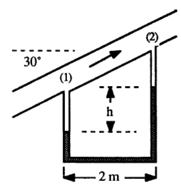
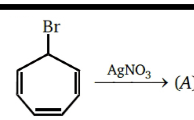
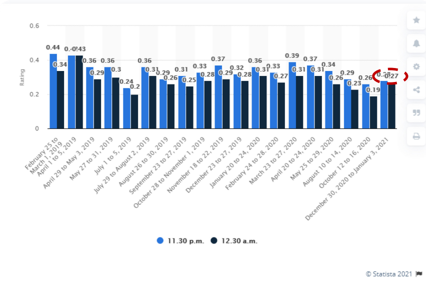


Fig. P2.35

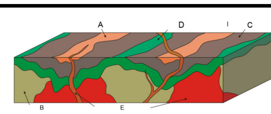
Q: Water flows upward in a pipe slanted at 30°, as in Fig. P2.35. The mercury manometer reads h = 12 cm. What is the pressure difference between points (1) and (2) in the pipe?
A. 12400 Pa
B. 36000 Pa
C. 26100 Pa
A: C



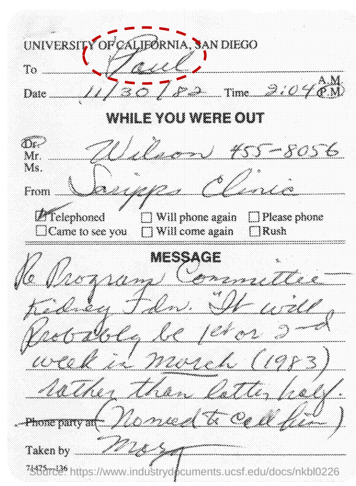
Q: Which statement is incorrect in respect of the above reaction ?
A. Product is aromatic
B. Product has high dipole moment
C. Product has less resonance energy
D. Product is soluble in polar solvent
A: C



Q: What was the rating of Adult Swim programming as of December 30, 2020?
A: 0.27



Q: What does the letter C represent?
A. Granite Intrusions
B. Granite-Gneiss
C. Mafic Volcanics
D. Ultramafic Volcanics
A: C


UNIVERSITY OF CALIFORNIA, SAN DIEGO

To Paul

Date 11/30/82 Time 2:04 A.M. P.M.

WHILE YOU WERE OUT

Dr. Mr. Ms. Wilson 455-8056

From Scripps Clinic

☑ Telephoned ☐ Will phone again ☐ Please phone
☐ Came to see you ☐ Will come again ☐ Rush

MESSAGE

Re Program Committee
Kidney Fdn. It will
probably be 1st or 2nd
week in March (1983)
rather than latter half.

Phone party at (Noted to call him)

Taken by Mary

Source: https://www.industrydocuments.ucsf.edu/docs/nkbl0226

Q: To whom is the document sent?
A: Paul

Figure D. More qualitative results on understanding tasks. Regions that related to the QAs are marked with red ellipses.

Cute winter dragon baby, kawaii, Pixar, ultra detailed, glacial background, extremely realistic

An ewok listening to music in headphones in a forest on a sunny day

Fantasy, a majestic sky filled with stars and galaxies, over looking a serene lake

Portrait with farmer and his black cow with horns .

Darth Maul as a knitted wool puppet

A beautiful photorealistic illustration of spring rain in colorful dark and deep tones.

A phoenix soaring above a city, aglow with golden flames

A detailed high resolution photograph of a captivating cyberpunk girl with vibrant pink hair looking intently at the camera as she stands confidently in a bustling cyberpunk town. The lighting features neonlit streets casting a mix of cool blues and warm pinks, highlighting the girls features and reflecting off the wet pavement. The colors include a palette of bold pinks, blues, and purples, with contrasting dark shadows and bright neon highlights.

Truck, water color art

A translucent birthday cake shape traced by of light particles

Beautiful landscape photography, summer, Indonesia

A painting of two people walking together in the rain in the evening, in the style of color splash

Whole cyberpunk badger wearing a yellow jacket on a white background, cartoon style, cyberpunk

A young boy in an outfit with many different colorful designer items, with colorful hat and cool glasses in the style of fantastical

Super cute little tiger rendered in the style of pixar cartoon, full body, shiny and fluffy, bright big eyes, fluffy tail, sweet smile, energetic and playful, exaggerated facial expression

Figure E. More qualitative results on text-to-image generation tasks.

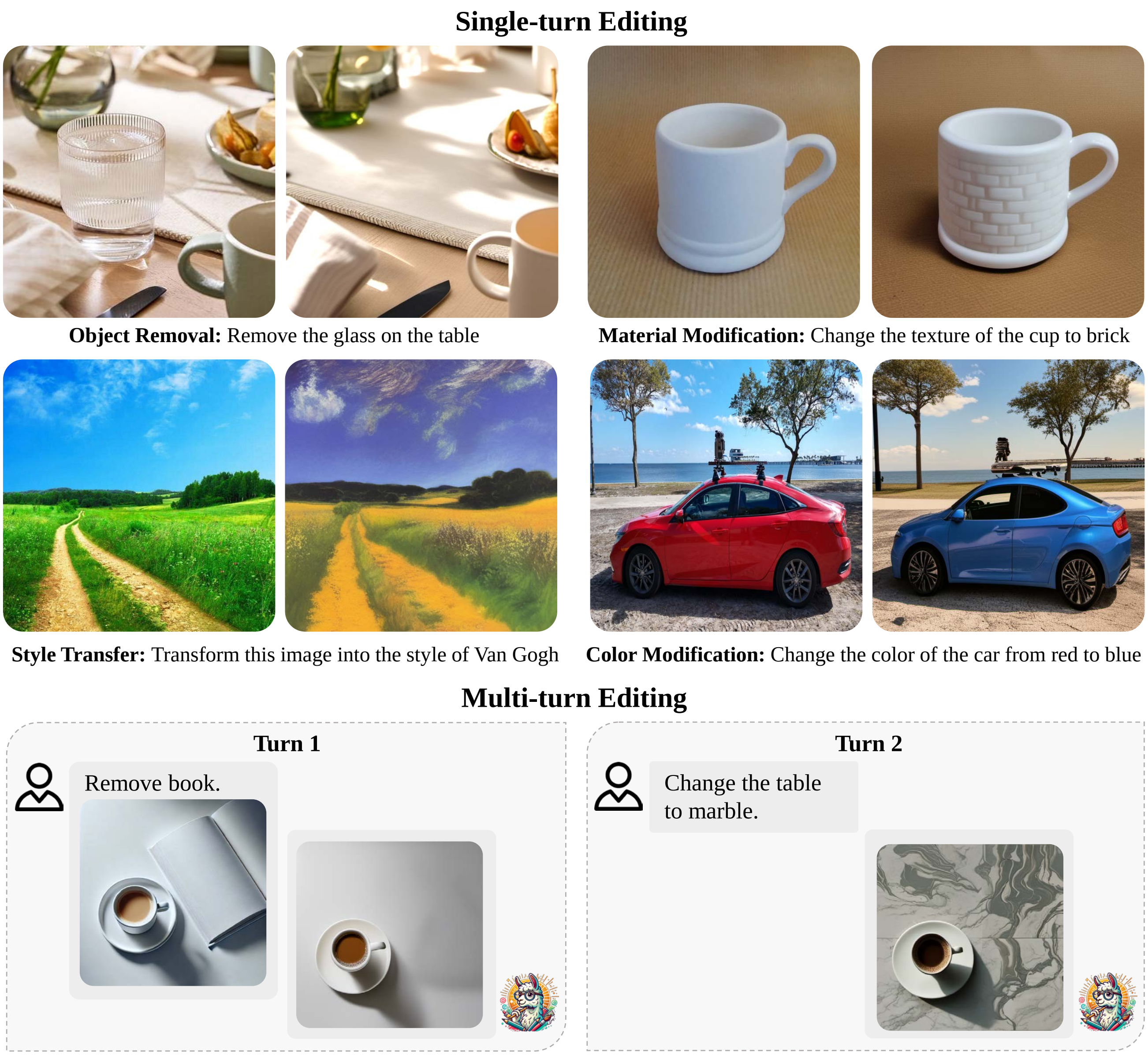


Figure F. More qualitative results on mixed-modal generation tasks.