Supplementary Material: ChatUMM

1. Detailed Data Pipeline

This section provides the detailed taxonomy, atomic operations, and stage-specific construction examples that are summarized in the main paper (Section 3).

1.1. Taxonomy of Conversational Data

To categorize our diverse conversational data, we introduce a formal four-dimensional taxonomy. Each data sample is classified based on the following core attributes:

User Input Modality: The type of input provided by the user in the final turn. This can be either text-only (`t`) or text and image (`ti`). Text-only (`t`) refers to commands issued entirely through text (e.g., "Make the sky blue"). In contrast, text-and-image (`ti`) involves providing both a textual instruction and a reference image. This is crucial for tasks like image editing, e.g., uploading a pet photo to ask, "Place my dog in a magical forest."

Model Output Modality: The type of output the model generates in the final turn. This can be either image-only (`i`) or text and image (`ti`). An image-only output (`i`) is the standard for most generative tasks, where the model's turn is to produce the requested visual content. In contrast, text-and-image (`ti`) enables continuous interaction by following the generated image with a relevant text response, integrating visual generation into the textual dialogue flow.

Historical Dependency Modality: The modality and quantity of historical turns the final turn depends on. This is classified as none (`0`), text history (`t1`, `tn`), or image history (`i1`, `in`). Zero dependency (`0`) denotes a context-free final turn. Text dependency (`t1`, `tn`) occurs when the final turn references prior textual dialogue (e.g., "Draw one for me" after discussing lions). Image dependency (`i1`, `in`) is crucial for stateful tasks like editing (e.g., "Make the hat red") or composition, requiring the final turn to retrieve and manipulate specific visual history.

Historical Dependency Depth: The number of turns separating the final turn from the historical context it relies on, classified as zero (`0`), immediate (`1`), or long-range (`n`). Zero dependency (`0`) signifies a context-free instruction. A depth of one (`1`) represents short-range dependency (e.g., editing the immediately preceding image). A depth of `n` ($n > 1$) represents long-range dependency, where the target context is separated by unrelated "distractor" turns. Training on this data is essential for robust context tracking, teaching the model to disregard unrelated intermediate information and resolve long-range dependencies.

This pipeline establishes a clear naming convention: `<input>_<output>_<dependency>_<depth>`. For instance, `t_i_i1_1` denotes a text input (`t`) generating an image (`i`), conditioned on one historical image (`i1`) from the immediately preceding turn (depth `1`).

1.2. LLM-powered Atomic Operations

To streamline the data synthesis pipeline, we design a suite of atomic, LLM-powered operations serving as modular building blocks for conversational logic. These operations accept structured text inputs (e.g., captions, queries) and synthesize conversational components, transforming single-turn data into fluid, context-aware dialogues. We categorize them into three groups corresponding to our pipeline stages: basic dialogue construction, long-range dependency injection, and interleaved output generation, as detailed in Table 1. Selected operations are visualized in Figure 1.

1.3. Stage (a): Basic Multi-turn Construction

This initial stage transforms standard single-turn datasets into elementary, stateful multi-turn dialogues. We source diverse high-quality data from text-to-image, image editing, and subject-driven generation tasks to construct dialogues with a dependency depth of `0` or `1`.

Single-Turn Text-to-Image (`t_i_0_0`): Using the `caption2query` operation, an image caption (e.g., "A golden retriever is running on the grass") is converted into a natural user request for image generation, forming a single-turn interaction with no historical dependency.

Basic Q&A-based Generation (`t_i_t1_1`): We employ `caption2QA_q` to construct a two-turn dialogue. An image caption generates a Q&A pair (e.g., "What are the features of Golden Retrievers?") serving as the first turn context. The subsequent request (e.g., "Create one for me") references this history (`t1`) at a dependency depth of `1`.

Table 1. **Overview of LLM-powered atomic operations.** These operations serve as modular building blocks within our data synthesis pipeline, each designed to perform specific text-based transformations (e.g., on captions, queries) to construct fluid, stateful dialogues. They are categorized by the pipeline stage in which they are primarily employed.

Atomic Operation	Description
Stage (a): Basic Multi-turn Construction	
`caption2query`	Converts an image caption into a user query for generating the image.
`caption2QA_q`	Transforms an image caption into a Q&A pair followed by a generic user query (e.g., "Create one for me") to initiate history-dependent generation.
`drive_hs`	Synthesizes a subject-driven query combining two subjects from the two immediately preceding turns (e.g., "Draw them together").
`drive_i_h`	Similar to `drive_hs`, but combines the subject in the uploaded image with a subject from the immediately preceding turn.
Stage (b): Independent Single-turn Insertion	
`query2dep_q`	Rewrites a user query (e.g., "Add a red hat") into a specific, explicit instruction (e.g., "Add a red hat to the dog that is reading a book") that resolves the reference to the image or subject to be edited.
`caption2QA_q_dep`	Transforms an image caption into a Q&A pair followed by a history-dependent user query (e.g., "Generate the dog we discussed earlier").
`drive_hs_dep`	Synthesizes a subject-driven query combining subjects from two prior turns separated from the current turn by "distractor" turns (e.g., "Draw the golden retriever lying on the table and the white cat lying on the open notebook together in the next image").
`drive_i_h_dep`	Similar to `drive_hs_dep`, but combines the subject in the uploaded image with a subject from a prior turn separated by "distractor" turns.
Stage (c): Interleaved Output Generation	
`Q_from_caption`	Generates a relevant general-knowledge question based on an image caption.
`A_from_caption`	Generates a factual answer to the question produced by `Q_from_caption`.

Transform the given image caption into **six** natural, instruction-like user requests for image generation.
Note:
1: Do not invent, guess, or describe anything not clearly present in the given image caption. Avoid speculative language (e.g., "seems," "maybe," "appears to," "like").
2: The output format should be **six** <Result><User></User></Result>. The language and length requirements for each Result are: (English, short), (English, medium), (English, long), (simplified Chinese, short), (simplified Chinese, medium), (simplified Chinese, long). Direct translation between English and simplified Chinese is prohibited. The two languages are independent of each other.
3: If there are text elements in the image caption, please keep the original language and do not translate them.
4: Use the following sentence structures:
English: {random.choice([Common Sentence Structures for Image Requests (English Version)])}
simplified Chinese: {random.choice([Common Sentence Structures for Image Requests (simplified Chinese Version)])}
The "..." in the reference sentence structures above needs to be replaced with meaningful content and should not be retained in the final generated user requests.

Now convert this image caption to **six** user requests (diversity matters): A lone individual sits on a weathered wooden bench on a grassy slope, facing a breathtaking alpine panorama. The person, seen from behind, wears a dark down jacket, grey trousers, and sneakers. Below the vantage point, a calm lake stretches out, flanked by a town with traditional alpine architecture and dense evergreen forests. The backdrop features towering mountains with rocky, snow-dusted faces, their peaks hidden by thick, swirling mist under an overcast sky.

<Result><User> Please display a lone person sitting on a bench on a grassy slope, overlooking a misty alpine landscape with a lake and mountains. </User></Result>
<Result><User> Please imagine a scene where an individual in a dark down jacket sits on a weathered wooden bench, facing a calm lake and a town with traditional alpine architecture surrounded by evergreen forests. </User></Result>
<Result><User> Please craft a detailed image of a person seen from behind, wearing grey trousers and sneakers, seated on a wooden bench. The figure should be looking out at a breathtaking panorama featuring a lake, dense forests, and towering, snow-dusted mountains with peaks obscured by thick, swirling mist under an overcast sky. </User></Result>

<Result><User> 给我看一张一个人坐在草坡上的旧木椅上眺望高山湖泊和云雾缭绕群山的照片。 </User></Result>
<Result><User> 给我展示一幅图像，画面中是一个穿着深色羽绒服的背影，正独自坐在风化的长椅上，俯瞰着下方宁静的湖泊以及被茂密常绿森林环绕的传统阿尔卑斯小镇。 </User></Result>
<Result><User> 给我一张图片，描绘一位穿着灰色裤子和运动鞋的游客坐在草坡上的木长椅上。在他的面前展现出壮丽的景色：平静的湖水延伸向远方，传统的城镇建筑错落有致，背景是岩石裸露且沾有积雪的巍峨大山，山顶在阴沉的天空下被厚重的迷雾所遮蔽。 </User></Result>

Modify the given user edit request so that it clearly specifies the previously generated image to be edited.
Assume there have been multiple rounds of unrelated conversation (possibly including generating images) between the previously generated image and the given user edit request. You must account for these unrelated conversations and accurately identify the specific image (i.e., the previously generated image) to be edited.
You will be provided with two types of information:
(1) caption: The caption of the previously generated image,
(2) query: The user edit request needed to be modified.
Note: The output format should be **six** <Result><User></User></Result>. The language and length requirements for each Result are: (English, short), (English, short), (English, short), (simplified Chinese, short), (simplified Chinese, short), (simplified Chinese, short). Direct translation between English and simplified Chinese is prohibited. The two languages are independent of each other.
Here are some examples:
Example 1:
caption: Rich still life features an overflowing wicker basket brimming with a vibrant assortment of fresh fruits like grapes, apples, and peaches, artfully arranged on a wooden table alongside a glass of red juice and scattered berries.
query: Change the background to a blue sky.
Reasonable Result:
<Result><User> Change the background of the wicker basket with fruits to a blue sky. </User></Result>
Example 2:
caption: A stylish young Asian woman with short dark hair, dressed in an elegant light grey two-piece outfit and carrying a crossbody bag, looking contemplatively to her side. She is situated in a modern industrial-style space characterized by grey concrete walls, dark metal structures, and geometric modern light fixtures, creating a minimalist yet sophisticated atmosphere.
query: Add a new coffee cup on the right table.
Reasonable Result:
<Result><User> For the image with the woman in an elegant light grey outfit, add a new coffee cup on the right table. </User></Result>
Example 3:
caption: A fair-skinned woman with shoulder-length dark hair and red lipstick stands slightly angled towards the camera. She wears a mauve ribbed top and long, dangling earrings. A framed mirror with a blue and gold ornate frame hangs on the wall behind her. The wall is a textured, reddish-pink color.
query: The woman tilts her head slightly to her right, her smile fades to a neutral expression, and her gaze shifts slightly upward.
Reasonable Result:
<Result><User> The woman wears a mauve ribbed long-sleeved top, her head tilts slightly to her right, her smile gradually fades to a neutral expression, and her gaze shifts slightly upward. </User></Result>

The given caption of the previously generated image: A young Asian man with short black hair and glasses stands in profile, wearing a grey jacket and carrying a black backpack. Behind him are buildings with wooden siding and beige stucco, while steep, snow-dusted mountains rise into a grey, overcast sky.
The given user edit request needed to be modified: Change the weather to sunny and the schoolbag to purple.
Modify the given user edit request (diversity matters)

<Result><User> Change the weather to sunny and the backpack to purple for the image of the young Asian man with glasses. </User></Result>
<Result><User> For the picture featuring a man in a grey jacket against snowy mountains, make the sky sunny and the bag purple. </User></Result>
<Result><User> Take the photo of the man standing in profile and change the overcast sky to sunny and his black backpack to purple. </User></Result>
<Result><User> 将那位背着黑色双肩包的戴眼镜男士所在的场景改为晴天，并将包变成紫色。 </User></Result>
<Result><User> 把那张背景是雪山和木屋的图片改成晴朗天气，同时把男生的书包改成紫色。 </User></Result>
<Result><User> 针对那个穿灰色夹克的亚洲男子的图片，把阴天背景换成大晴天，背包颜色换成紫色。 </User></Result>

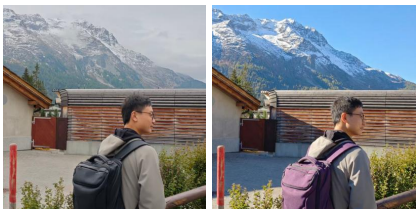


Figure 1. **Visualization of selected atomic operations. Top:** `caption2query` converts image captions into user queries. **Bottom:** `query2dep_q` transforms user queries into specific, explicit instructions that clearly identify the target image or subject.

Table 2. **Comparison with state-of-the-arts on visual understanding benchmarks.** MME-S refers to the summarization of MME-P and MME-C. For MoE models, we report their activated params / total params. †: MetaQuery [19] adopts the pre-trained model from Qwen2.5-VL [1] and freezes it during training. $**$: Partial results are from by MetaMorph [25] or MetaQuery [19]. $*$: We report the results without Chain-of-Thought.

Type	Model	# LLM Params	MME-P↑	MME-S↑	MMBench↑	MMMU↑	MM-Vet↑	MathVista↑	MMVP↑
Und. Only	InternVL2 [9]	1.8B	1440	1877	73.2	34.3	44.6	46.4	35.3
	InternVL2.5 [8]	1.8B	-	2138	74.7	43.6	60.8	51.3	-
	Qwen2-VL [27]	1.5B	-	1872	74.9	41.1	49.5	43.0	-
	Qwen2.5-VL [1]	3B	-	2157	79.1	53.1	61.8	62.3	-
	BLIP-3 [35]	4B	-	-	76.8	41.1	-	39.6	-
	LLava-OV [16]	7B	1580	-	80.8	48.8	57.5	63.2	-
	InternVL2 [9]	7B	1648	2210	81.7	49.3	54.2	58.3	51.3
	InternVL2.5 [8]	7B	-	2344	84.6	56.0	62.8	64.4	-
	Qwen2-VL [27]	7B	-	2327	83.0	54.1	62.0	58.2	-
	Qwen2.5-VL [1]	7B	-	2347	83.5	**58.6**	67.1	68.2	-
	Emu3-Chat** [28]	8B	1244	-	58.5	31.6	37.2	-	36.6
	Kimi-VL [15]	2.8B/16B	-	-	-	57.0	66.7	68.7	-
	DeepSeek-VL2 [31]	4.1B/28B	-	-	-	51.1	60.0	62.8	-
Unified	Show-o$_{512}$ [33]	1.3B	1097	-	-	26.7	-	-	-
	Janus [29]	1.5B	1338	-	69.4	30.5	34.3	-	-
	Janus-Pro [7]	1.5B	1444	-	75.5	36.3	39.8	-	-
	BAGEL [10]	1.5B MoT	1610	2183	79.2	43.2	48.2	63.4	54.7
	ILLUME [26]	7B	1445	-	75.1	38.2	37.0	-	-
	VILA-U$^{**}_{256}$ [30]	7B	1336	-	66.6	32.2	27.7	-	22.0
	Chameleon** [5]	7B	-	-	35.7	28.4	8.3	-	0.0
	Janus-Pro [7]	7B	1567	-	79.2	41.0	50.0	-	-
	MetaQuery-XL† [19]	7B	1685	-	83.5	**58.6**	66.6	-	-
	LlamaFusion** [24]	8B	1604	-	72.1	41.7	-	-	-
	MetaMorph [25]	8B	-	-	75.2	41.8	-	-	48.3
	SEED-X [12]	13B	1457	-	70.1	35.6	43.0	-	-
	TokenFlow-XL [21]	13B	1546	-	68.9	38.7	40.7	-	-
	MUSE-VL [34]	32B	-	-	81.8	50.1	-	55.9	-
	BAGEL [10]	7B MoT	**1687**	**2388**	**85.0**	53.7*	**67.2**	73.1	**69.3**
	ChatUMM (Ours)	7B MoT	1685	2382	84.6	53.8*	66.4	**74.7**	68.7

Single-Turn Image Editing (`ti_i_0_0`): This is a standard editing task involving a text query and a reference image, resulting in an modified image with no historical dependency.

Sequential Image Editing (`t_i_i1_1`): A standard single-step editing task is split into a two-turn dialogue. First, the user requests the original image (e.g., "Can you give me a photo of a golden retriever? I want to see it reading a book"). In the second turn, a targeted edit (e.g., "Add a red hat") is issued. The model must retrieve the image generated in the first turn (`i1`) to execute the edit, creating a dependency at a depth of `1`.

Basic Subject-Driven Generation (`t_i_in_1` and `ti_i_i1_1`): For subject composition, we synthesize multi-step dialogues. In `t_i_in_1`, we use `caption2query` twice to generate distinct subjects in consecutive turns (e.g., a white cat, then a golden retriever), followed by `drive_hs` to issue a composition command (e.g., "Draw them together"). Alternatively, for `ti_i_i1_1`, we generate a single subject (e.g., a white cat) via `caption2query`, then employ `drive_i_h` to combine it with a newly uploaded image (e.g., "Put them together").

1.4. Stage (b): Independent Single-turn Insertion

To equip the model with robust context tracking capabilities across noisy histories, we augment the dialogues from Stage (a) by inserting "distractor" turns (text-to-image, image understanding, or text chat). The core innovation is applying operations like `query2dep_q` to transform the user query into a history-dependent instruction. This process elevates the dependency depth from `1` to `n`, forcing the model to filter out distractors and resolve long-range dependencies.

Long-Context Q&A-based Generation (`t_i_t1_n`): Originating from a `t_i_t1_1` sample, the implicit request (e.g., "Create one for me") becomes ambiguous after distractor insertion. We employ `caption2QA_q_dep` to rewrite it into a specific instruction (e.g., "Generate the dog we discussed earlier") that explicitly references the textual topic from `n` turns ago.

Table 3. **Evaluation of text-to-image generation ability on GenEval benchmark.** 'Gen. Only' stands for an image generation model, and 'Unified' denotes a model that has both understanding and generation capabilities. † refers to the methods using the LLM rewriter.

Type	Model	Single Obj.	Two Obj.	Counting	Colors	Position	Color Attri.	Overall↑
Gen. Only	PixArt-α [6]	0.98	0.50	0.44	0.80	0.08	0.07	0.48
	SDv2.1 [23]	0.98	0.51	0.44	0.85	0.07	0.17	0.50
	DALL-E 2 [22]	0.94	0.66	0.49	0.77	0.10	0.19	0.52
	Emu3-Gen [28]	0.98	0.71	0.34	0.81	0.17	0.21	0.54
	SDXL [20]	0.98	0.74	0.39	0.85	0.15	0.23	0.55
	DALL-E 3 [2]	0.96	0.87	0.47	0.83	0.43	0.45	0.67
	SD3-Medium [11]	0.99	0.94	0.72	0.89	0.33	0.60	0.74
	FLUX.1-dev† [3]	0.98	0.93	0.75	0.93	0.68	0.65	0.82
Unified	Chameleon [5]	-	-	-	-	-	-	0.39
	LWM [17]	0.93	0.41	0.46	0.79	0.09	0.15	0.47
	SEED-X [12]	0.97	0.58	0.26	0.80	0.19	0.14	0.49
	TokenFlow-XL [21]	0.95	0.60	0.41	0.81	0.16	0.24	0.55
	ILLUME [26]	0.99	0.86	0.45	0.71	0.39	0.28	0.61
	Janus [29]	0.97	0.68	0.30	0.84	0.46	0.42	0.61
	Transfusion [38]	-	-	-	-	-	-	0.63
	Emu3-Gen† [28]	0.99	0.81	0.42	0.80	0.49	0.45	0.66
	Show-o [33]	0.98	0.80	0.66	0.84	0.31	0.50	0.68
	Janus-Pro-7B [7]	0.99	0.89	0.59	0.90	0.79	0.66	0.80
	MetaQuery-XL† [19]	-	-	-	-	-	-	0.80
	BAGEL† [10]	0.98	0.95	0.84	0.95	0.78	0.77	**0.88**
	ChatUMM (Ours)†	0.99	0.96	0.79	0.92	0.69	0.77	0.85

Table 4. **Comparison on GEdit-Bench.**

Type	Model	GEdit-Bench-EN (Full set)↑			GEdit-Bench-CN (Full set)↑		
		G_SC	G_PQ	G_O	G_SC	G_PQ	G_O
Private	Gemini 2.0 [13]	6.73	6.61	6.32	5.43	6.78	5.36
	GPT-4o [14]	**7.85**	**7.62**	**7.53**	**7.67**	**7.56**	**7.30**
Open-source	Instruct-Pix2Pix [4]	3.58	5.49	3.68	-	-	-
	MagicBrush [37]	4.68	5.66	4.52	-	-	-
	AnyEdit [36]	3.18	5.82	3.21	-	-	-
	OmniGen [32]	5.96	5.89	5.06	-	-	-
	Step1X-Edit [18]	7.09	6.76	6.70	7.20	6.87	6.86
	BAGEL [10]	7.36	6.83	6.52	7.34	6.85	6.50
	ChatUMM (Ours)	7.69	7.21	6.95	7.40	7.24	6.73

Long-Context Image Editing (`t_i_i1_n`): We insert distractors between the initial image generation and the editing request in a `t_i_i1_1` dialogue. Using `query2dep_q`, the ambiguous query (e.g., "Add a red hat") is transformed into a specific, explicit instruction (e.g., "Add a red hat to the dog that is reading a book"). This forces the model to ignore intermediate distractors and accurately retrieve the specific visual context from a depth of `n`.

Long-Context Subject-Driven Generation (`t_i_in_n` and `ti_i_i1_n`): For subject composition, we insert distractors before the final composition request. For `t_i_in_n`, `drive_hs_dep` expands a vague command (e.g., "Draw them together") into a history-dependent query (e.g., "Draw the golden retriever lying on the table and the white cat lying on the open notebook together in the next image"), explicitly combining subjects from prior turns separated by distractors. Similarly, for `ti_i_i1_n`, `drive_i_h_dep` combines a historical generated image with a user upload (e.g., "Draw the white cat lying on the open book together with this dog I uploaded").

1.5. Stage (c): Interleaved Output Generation

The final stage evolves the model's output from a single image (`i`) to an interleaved text-and-image (`ti`) format. Universally applied to the final turn of dialogues from Stages (a) and (b), we leverage operations like `Q_from_caption` and `A_from_caption` to generate a Q&A pair derived from the image of the final turn. The user's final instruction (`<query-final>`) is augmented with the generated question (`<query-final><Q>`), training the model to produce a interleaved response: the requested image followed by the textual answer (`<image><A>`). This stage updates task signatures (e.g., `t_i_0_0` → `t_ti_0_0`), integrating visual generation into the textual dialogue flow.

2. Extended Results and Qualitative Examples

This section details extended results and qualitative examples. Table 2, Table 3, and Table 4 expand on comparisons for visual understanding, text-to-image generation, and instruction-guided editing, respectively. Figure 2 and Figure 3 illustrate multi-turn cases, ranging from basic interactions to complex, distractor-robust scenarios. These evaluations confirm that our model achieves state-of-the-art performance in visual understanding and editing, maintains high fidelity in text-to-image generation, and exhibits superior robustness in complex dialogue scenarios.

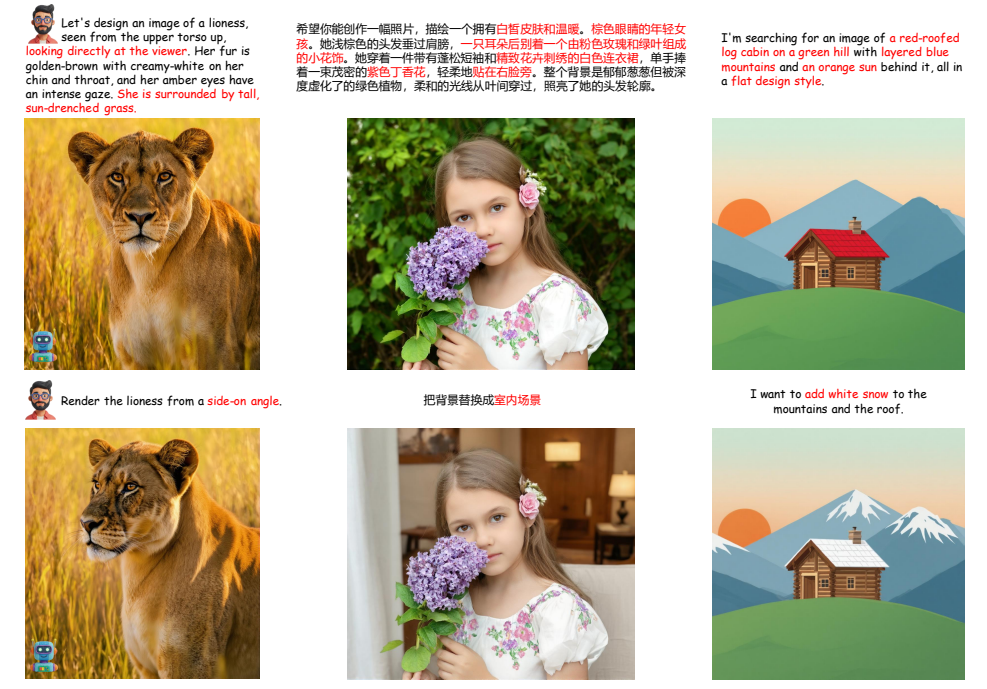


Figure 2. **Basic multi-turn interactions. Top:** The model ensures subject consistency during editing. **Middle:** Context-aware generation from text history and merging prior subjects into a single scene. **Bottom:** Interleaved sequence of knowledge retrieval, generation, editing.

第一张图：一位骑士。细节：全身板甲（银色），手握长剑，眼神坚毅，写实主义风格，环境是城堡一角。

第二张图：一位公主。细节：丝绸连衣裙（粉色），头戴钻石皇冠，金色卷发，童话风格，环境是皇家花园。

第三张图：把骑士转换成16位像素艺术风格。

第四张图：把骑士变成骷髅战士。

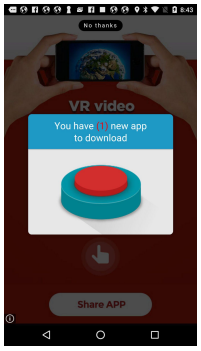
第五张图：把骑士的背景改成夜晚的城堡和月光。

The goal is to generate an image of a woman from the waist up, slightly off-center. She has dark brown straight hair with bangs, wears a light grey face mask and a beige cardigan, and her head is tilted back sharply, gazing upwards.

I'm looking for an image of a young East Asian man with short, spiky black hair and a dark grey zip-up hoodie. He's standing on train tracks under a large metal truss bridge, with warm sunset light illuminating his right side.

Can you render a scene featuring a mummy skeleton with a wood-grained texture, partially wrapped in frayed bandages, sitting upright in a dark stone sarcophagus? Its right hand is raised towards a light beam, surrounded by a flurry of pastel chalk pieces, and the sarcophagus rim is smudged with colorful chalk dust.

Give me an image of a man with his hands in his pockets, standing on a reflective blue floor and looking out a large window. The view is an urban landscape under a dramatic orange and red sunset, all rendered with the vibrant, textured brushstrokes of Fauvism.



How many new applications are there to download?

There is 1 new application to download.

I'd like to see the young man who was standing on the train tracks in front of the truss bridge, and the woman with straight hair and bangs who was wearing a beige V-neck cardigan. Place them together in an orchard during a harvest festival. The man should be guiding the woman through a VR device, and they should be surrounded by swirling holograms.
What are the functions of VR devices?

VR (Virtual Reality) devices have a variety of functions, primarily centered around creating immersive, simulated environments. Key functions include:
1. **Immersive Entertainment:** This is the most popular function. VR provides an immersive gaming experience, allowing players to feel physically present in the game world. It's also used for watching 360-degree videos and movies in a virtual cinema.
2. **Education and Training:** VR allows for realistic and safe simulations. Pilots can practice flying, surgeons can rehearse complex operations, and technicians can learn to repair machinery without real-world risks or high costs.
3. **Social Interaction:** Social VR platforms (like VRChat or Meta Horizon Worlds) enable users to create avatars, interact with others in virtual spaces, attend events, and play games together, regardless of their physical location.
4. **Design and Visualization:** Architects, engineers, and designers use VR to create and explore virtual prototypes of buildings, cars, or products. This allows them to walk through a design and make adjustments before any physical construction begins.
5. **Healthcare and Therapy:** In medicine, VR is used for pain management, physical rehabilitation, and treating psychological disorders like PTSD, anxiety, and phobias through controlled exposure therapy.
6. **Virtual Tourism and Exploration:** Users can visit famous landmarks, explore museums, or even travel to distant planets from the comfort of their homes, making travel accessible to everyone.

Figure 3. **Advanced multi-turn dialogues. Left:** Targeted subject editing. The model correctly edits the "Knight" from the first turn, effectively bypassing the unrelated "Princess" generated in the second turn to perform history-dependent editing. **Right:** Long-range subject-driven generation. The model combines two characters from the initial turns, effectively filtering out a sequence of "distractor" turns (e.g., Mummy, Fauvism painting) to execute the final composition.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 4

[2] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023. 5

[3] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 5

[4] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, pages 18392–18402, 2023. 5

[5] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024. 4, 5

[6] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In *ECCV*, pages 74–91, 2024. 5

[7] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025. 4, 5

[8] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 4

[9] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12):220101, 2024. 4

[10] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025. 4, 5

[11] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024. 5

[12] Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint arXiv:2404.14396*, 2024. 4, 5

[13] Google Gemini. Experiment with gemini 2.0 flash native image generation. https://developers.googleblog.com/en/experiment-with-gemini-20-flash-native-image-generation/, 2025. 5

[14] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 5

[15] Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025. 4

[16] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 4

[17] Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with blockwise ringattention. *arXiv preprint arXiv:2402.08268*, 2024. 5

[18] Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025. 5

[19] Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, et al. Transfer between modalities with metaqueries. *arXiv preprint arXiv:2504.06256*, 2025. 4, 5

[20] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 5

[21] Liao Qu, Huichao Zhang, Yiheng Liu, Xu Wang, Yi Jiang, Yiming Gao, Hu Ye, Daniel K Du, Zehuan Yuan, and Xinglong Wu. Tokenflow: Unified image tokenizer for multimodal understanding and generation. In *CVPR*, pages 2545–2555, 2025. 4, 5

[22] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 5

[23] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 5

[24] Weijia Shi, Xiaochuang Han, Chunting Zhou, Weixin Liang, Xi Victoria Lin, Luke Zettlemoyer, and Lili Yu. Lmfusion: Adapting pretrained language models for multimodal generation. *arXiv preprint arXiv:2412.15188*, 2024. 4

[25] Shengbang Tong, David Fan, Jiachen Zhu, Yunyang Xiong, Xinlei Chen, Koustuv Sinha, Michael Rabbat, Yann LeCun, Saining Xie, and Zhuang Liu. Metamorph: Multimodal understanding and generation via instruction tuning. *arXiv preprint arXiv:2412.14164*, 2024. 4

[26] Chunwei Wang, Guansong Lu, Junwei Yang, Runhui Huang, Jianhua Han, Lu Hou, Wei Zhang, and Hang Xu. Illume: Illuminating your llms to see, draw, and self-enhance. *arXiv preprint arXiv:2412.06673*, 2024. 4, 5

[27] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin

Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 4

[28] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024. 4, 5

[29] Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *CVPR*, pages 12966–12977, 2025. 4, 5

[30] Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. Vila-u: a unified foundation model integrating visual understanding and generation. *arXiv preprint arXiv:2409.04429*, 2024. 4

[31] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024. 4

[32] Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. In *CVPR*, pages 13294–13304, 2025. 5

[33] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024. 4, 5

[34] Rongchang Xie, Chen Du, Ping Song, and Chang Liu. Muse-vl: Modeling unified vlm through semantic discrete encoding. *arXiv preprint arXiv:2411.17762*, 2024. 4

[35] Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S Ryoo, et al. xgen-mm (blip-3): A family of open large multimodal models. *arXiv preprint arXiv:2408.08872*, 2024. 4

[36] Qifan Yu, Wei Chow, Zhongqi Yue, Kaihang Pan, Yang Wu, Xiaoyang Wan, Juncheng Li, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. Anyedit: Mastering unified high-quality image editing for any idea. In *CVPR*, pages 26125–26135, 2025. 5

[37] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *NeurIPS*, pages 31428–31449, 2023. 5

[38] Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024. 5