QWEN-IMAGE-OMNI: COMPLEX VISUAL PROGRAMS DON'T NEED EXTERNAL REASONER

Anonymous authors
Paper under double-blind review

ABSTRACT

Unified multimodal models (UMMs) support multi-turn understanding and generation of interleaved images and text. Complex visual programming requires finer control at each turn: choosing an operation, identifying its target, and recalling historical images with distinct roles while preserving earlier decisions. We present Qwen-Image-Omni, which connects a pretrained Vision-Language Model (VLM) and Diffusion Transformer (DiT) through structured skill calls and a shared visual history. The VLM uses its internal reasoning to select the next operation, assign reference roles, and expand the goal into an executable visual instruction. The same VLM supplies the DiT's conditioning, and each generated image becomes available for subsequent reasoning and execution. We construct histories that preserve dependencies between single-turn skills, then adapt visual execution and prompt expansion in two stages. Autoregressive supervision targets visual prompt content while system instructions specify call syntax. Under matched training and call budgets, the combined skill-and-reference interface improves complex visual design and historical editing over general instructions with natural-language references. Public benchmarks further demonstrate strong interleaved generation, reasoning-informed editing, and general generation and editing capability.

1 INTRODUCTION

Unified multimodal models (UMMs) connect image understanding and generation (Sun et al., 2024b; Ge et al., 2024a; Chameleon Team, 2024; Wu et al., 2025b), increasingly supporting interleaving and editing (Zhou et al., 2025; Xie et al., 2025a; Deng et al., 2025; Cui et al., 2025). Complex visual programming places additional demands on this interaction. Each step must identify an operation, distinguish its editing target from identity or layout references, and preserve decisions that should survive the edit. Later steps may reuse an early input alongside a recent result. The challenge is to make these choices available to image execution throughout a program.

We introduce **Qwen-Image-Omni**, a UMM that expresses these choices as structured skill calls over a shared image-text history. A skill tag identifies the intended transformation, reference IDs identify its inputs, and the visual prompt assigns their roles and specifies what to preserve. Together they condition the DiT: the same images can support clothing transfer, subject composition, or style transfer, each requiring different changes and preservation constraints. The VLM constructs the call and supplies its continuous conditioning. Generated images return to the shared history with stable IDs for later reuse.

This gives *internal reasoning* a concrete role: choosing the next operation, assigning historical-image roles, and expressing the required visual change. Figure 1 illustrates poster creation that returns to an original sketch, historical editing that combines identity, clothing, pose, and scene, and coherent text–image generation. Programs can follow a specified operation order or construct calls from a broader goal.

Our adaptation addresses both execution and instruction construction. Dependency-preserving mixing turns single-turn generation and editing examples into histories with reusable references. The first stage trains the DiT to execute skill-conditioned instructions in these histories. The second adapts the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

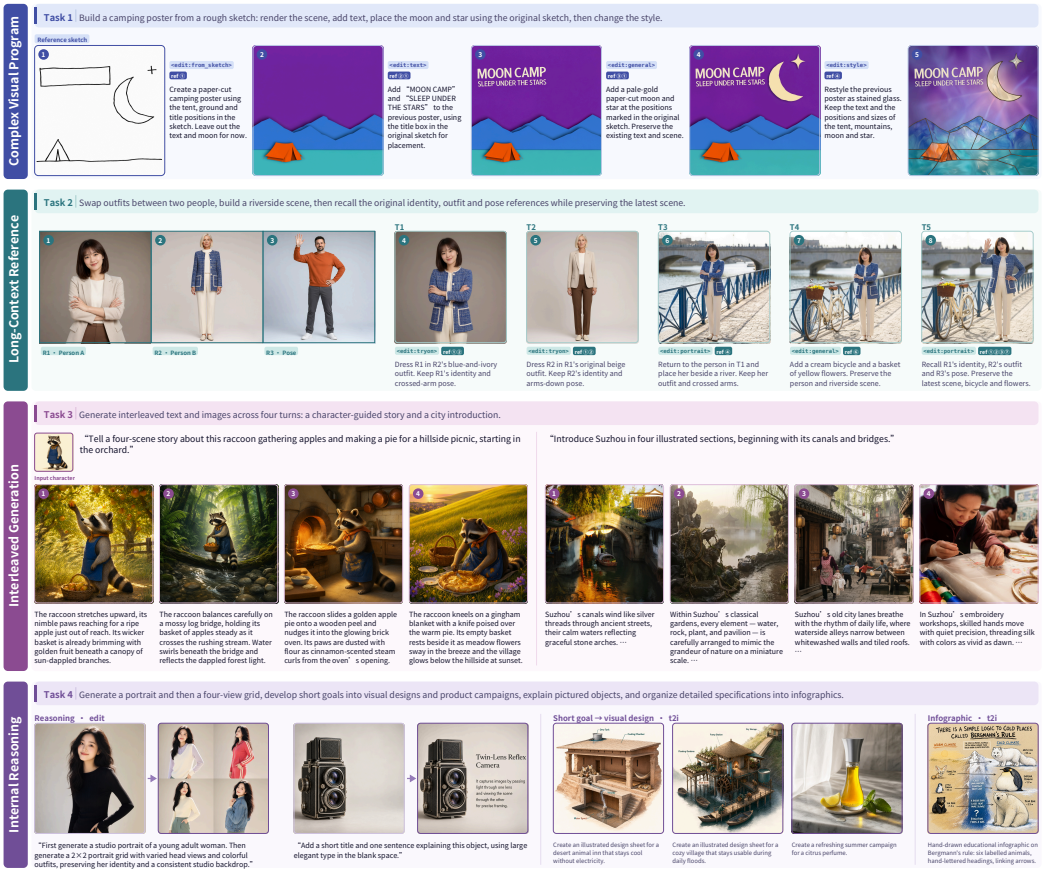

Figure 1: Qwen-Image-Omni composes visual operations and reuses their results within a shared history. **(1) Complex visual programming:** a camping poster evolves through sketch rendering, text insertion, object placement, and restyling. **(2) Role-aware historical reference use:** the final edit combines R1's identity, R2's outfit, R3's pose, and the latest riverside scene, retaining the bicycle and flowers. **(3) Coherent interleaved generation:** four-turn stories and city introductions coordinate text and images. **Bottom: Internal reasoning** illustrates portrait editing, object explanation, goal expansion, and infographic generation.

VLM's visual prompt expansion, with autoregressive supervision confined to `<prompt>` content and flow matching maintaining compatible visual conditioning. Call syntax remains specified through system instructions. Our contributions are:

1. A native visual-program interface that conditions image execution on operation types, reference roles, and preservation constraints, with shared history for reusing intermediate images.
2. A multi-turn adaptation strategy that constructs dependency-preserving histories from single-turn skills and trains visual execution and prompt expansion in two stages.
3. Matched interface comparisons supporting complex design and historical editing, with strong public-benchmark results in generation, editing, and interleaving.

2 Related Work

Unified Multimodal Models. UMMs connect language modeling, visual representations, and decoders (Koh et al., 2023; Sun et al., 2024b; Ge et al., 2024a; Dong et al., 2024; Sun et al., 2024a; Ge et al., 2024b; Tian et al., 2024; Tong et al., 2025), using discrete-token or hybrid autoregressive–diffusion architectures (Chameleon Team, 2024; Wang et al., 2026c; Wu et al., 2025c;b; Chen et al.,

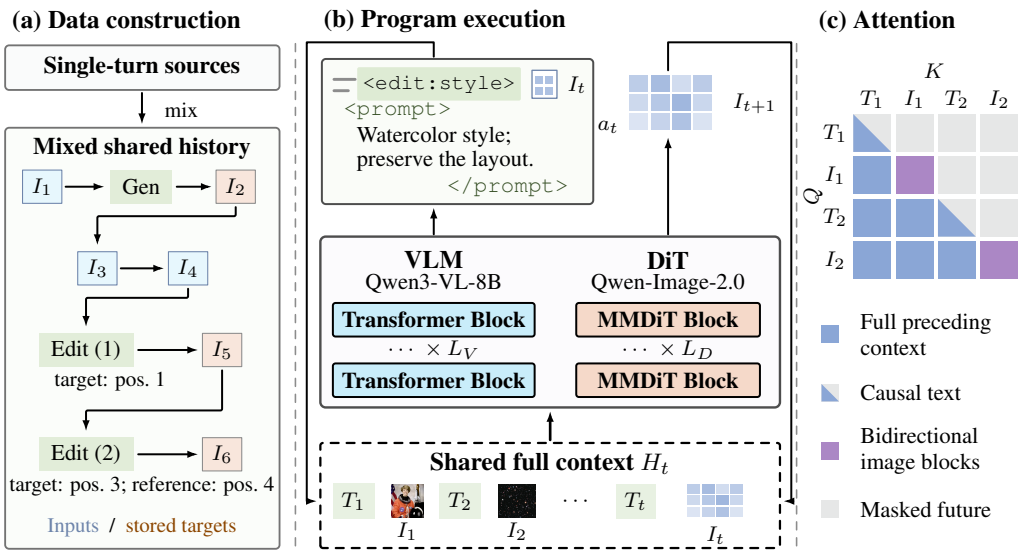

Figure 2: Training and executing visual programs over a shared history. (a) Dependency-preserving mixing places unrelated tasks between inputs and edits, retaining addressable inputs and stored targets. (b) The VLM constructs a skill call and the DiT executes it; both outputs enter the history for later operations. Generated latents remain with the DiT; decoded images enter the VLM. (c) Causal text and block-causal images retain preceding context. Conditioning and codec paths are omitted. Photographs: NASA; calls and outputs are schematic.

2025b; Zhou et al., 2025; Kou et al., 2025; Xie et al., 2025a; Ma et al., 2025; Yang et al., 2025a). BAGEL, Emu3.5, Show-o2, and SenseNova-U1/U1.5 extend native interleaving (Deng et al., 2025; Cui et al., 2025; Xie et al., 2025b; Diao et al., 2026b;a).

Historical Context. CoMM and ILLUME-X develop coherent interleaving (Chen et al., 2025a; Wang et al., 2026a). Wan-Weaver uses dense prompt context and VAE references (Xing et al., 2026); DuoGen conditions a video DiT on prior multimodal features and images (Shi et al., 2026). WEAVE and Edit-R2 study historical recall and multi-turn editing (Chow et al., 2026; Ye et al., 2026). Our method builds on historical conditioning to express each step as a skill call: the prompt assigns targets, reference roles, and preservation constraints, and the output remains addressable by later calls. This operation-level interface connects historical context to the execution of a visual program.

Skills and Reasoning. Latent diffusion and DiTs (Rombach et al., 2022; Peebles & Xie, 2023) support spatial control and instruction editing (Zhang et al., 2023; Brooks et al., 2023; Yu et al., 2025), with broader interfaces in OmniGen, Kontext, and Qwen-Image (Xiao et al., 2025; Wu et al., 2026; Black Forest Labs et al., 2025; Wu et al., 2025a; Yin et al., 2026a). Visual programs invoke modules (Gupta & Kembhavi, 2023; Surís et al., 2023; Shen et al., 2023); reasoning also guides instruction expansion and refinement (Fu et al., 2024; Huang et al., 2024; Wang et al., 2026b; Bian et al., 2026). We adapt pretrained reasoning to compose skills with historical references and preservation constraints.

3 Qwen-Image-Omni: Native Visual Program Execution

Our method combines addressable visual operations, histories constructed from single-turn skills, and two-stage adaptation of execution and prompt expansion (Figures 2 and 3). The pretrained VLM, DiT, feature projection, and image codec are inherited from the base model.

3.1 Model Architecture and Visual Programs

Shared history. We build on Qwen-Image-2.0 (Zhao et al., 2026), with Qwen3-VL-8B (Bai et al., 2025) as VLM/condition encoder and MMDiT (Esser et al., 2024) as generator. Both components access the user's request and complete image–text history H_t through their respective embeddings

(Figure 2): the VLM reasons about the next action, while the DiT resolves its references and synthesizes the result. Given call a_t, the DiT generates z_{t+1} conditioned on historical latents $Z(H_t)$ and VLM features $c_\theta(H_t, a_t)$. VAE decoding gives I_{t+1} and $H_{t+1} = [H_t, a_t, I_{t+1}]$, making each output reusable as an edit target or reference.

Conditioning and learned addressing. The inherited projection maps final-layer VLM features for history and call into the DiT text stream. Its visual stream concatenates clean historical latents with the noisy target. Image IDs occupy the temporal positional coordinate; spatial coordinates restart per block. Text attends to preceding content; image tokens attend bidirectionally within their block and to earlier blocks, while future blocks are masked. No additional task-dependent attention mask selects references. The DiT learns to locate historical images and bind their roles within this complete context.

Native visual state. External images enter the DiT through the base VAE encoder (Kingma & Welling, 2014; Rombach et al., 2022). Generated latents remain in its visual context, avoiding another VAE decode-and-re-encode cycle. Decoded images enter the VLM's image encoder. The VLM receives each actual result before selecting the next operation, supplying both reasoning and continuous conditioning.

Operation-level control. A call $a_t = (s_t, R_t, p_t)$ specifies a skill, an ordered list of historical image IDs, and a visual instruction. The tuple is a notational shorthand: calls are expressed as structured natural-language instructions, with tags identifying the fields. The skill conditions the intended transformation; references identify its inputs; the instruction assigns their roles and preservation constraints. For example, clothing transfer should modify the target's outfit while retaining its identity, whereas composition may combine subjects. These skills share a DiT: tags provide semantic conditioning, not separate experts or hard attention gates. A generic tag leaves the transformation to the remaining instruction and context. The VLM reasons over user requests and history, expands visual instructions, and supplies continuous features for execution.

Reference addressing. Input and generated images receive permanent global IDs. Generation calls contain a skill tag and `<prompt>` content without a reference block; editing adds a nonempty `<ref>` list of `<image N>` entries. In our goal-driven protocol, image numbers inside the instruction denote positions in this ordered list. With $R_t = (7, 2, 3)$, "the first image" means global image 7; images 2 and 3 can supply clothing and pose while image 7 remains the target. A single reference is described as "the image." IDs identify images, while the prompt assigns their roles. Every prior image remains visible.

A worked visual program. Figure 1's camping-poster sketch has ID 1; successive outputs have IDs 2–5. The four calls take the following form:

1. `<edit:from_sketch><ref><image 1></ref><prompt>`Render the sketch as a paper-cut camping scene; retain layout and omit text and moon.`</prompt></edit:from_sketch>`
2. `<edit:text><ref><image 2><image 1></ref><prompt>`Add "MOON CAMP" and "SLEEP UNDER THE STARS" to the first image using the second image's title placement; preserve the scene.`</prompt></edit:text>`
3. `<edit:general><ref><image 3><image 1></ref><prompt>`Add a moon and star to the first image at the second image's marked positions; preserve text and scene.`</prompt></edit:general>`
4. `<edit:style><ref><image 4></ref><prompt>`Restyle the image as stained glass; preserve text, object positions, and sizes.`</prompt></edit:style>`

The sketch supplies layout; newer images supply the editable state. The user specifies the order; the model constructs calls, assigns references, and expands instructions.

3.2 From Single-Turn Skills to Multi-Turn Training

Skill-structured sequences. We use 10 million generation and 5 million editing examples across nearly 30 skills; target images come from existing caption–image pairs. Sampled examples are interleaved subject to input-before-target dependencies; independent tasks may separate references from edits. IDs are assigned after serialization and calls rewritten accordingly. Stored targets form

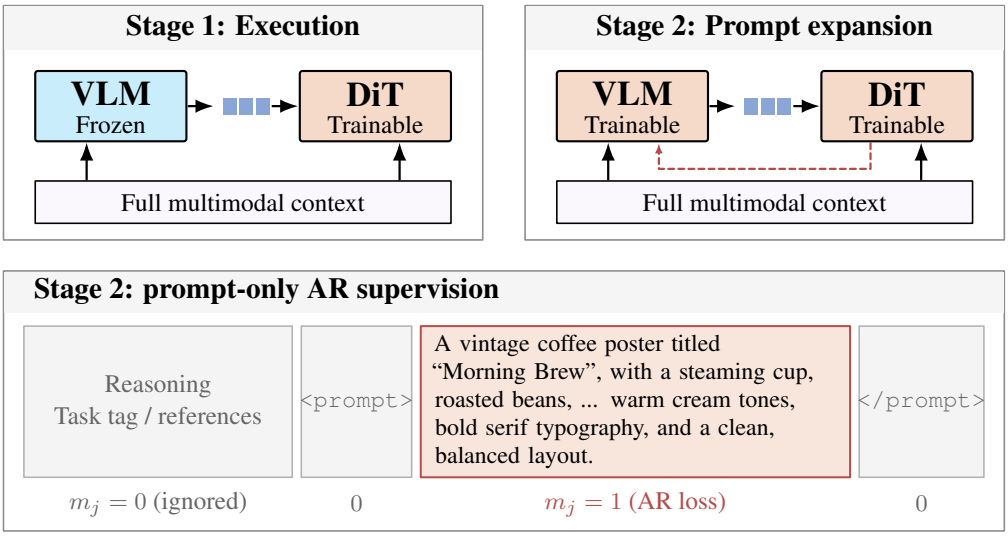

Figure 3: Adapting visual execution and instruction construction. Stage 1 trains the DiT to execute calls in shared histories. Stage 2 trains both models, with AR loss only on <prompt> content. Reasoning, tags, references, and delimiters remain in context without direct AR supervision; system prompts specify call syntax. Blue blocks are continuous conditioning features, and the dashed arrow carries FM gradients to the VLM.

training history, with up to eleven target rounds. This preserves each operation's supervision while varying its input positions (Figure 2a). Interleaving independent tasks separates edits from their references while preserving input–target dependencies.

For example, Figure 2a inserts images 3 and 4 before an edit targeting image 1; a later edit targets image 3 with image 4 as reference. The next target therefore cannot be identified from recency alone.

Coherent sequences. We supplement the mixed histories with native multi-turn image–text sequences, including comics, documents, and model-constructed examples, as well as editing sequences derived from sampled video frames. These sources provide coherent visual context alongside the mixed-skill histories. FM supervision applies to the current target image, while historical images remain clean.

3.3 Adapting Execution and Prompt Expansion

Stage 1: execution. We first adapt the visual executor to skill calls in shared histories while keeping the VLM fixed. The DiT is trained with a learning rate of 2×10^{-5}. For a clean target z, Gaussian noise ϵ, and $s \sim \mathcal{U}[0, 1]$, we use $z_s = (1-s)z + s\epsilon$ and target velocity $\epsilon - z$ (Lipman et al., 2023; Liu et al., 2023). The FM loss is mean squared velocity error over the current target's latent coordinates, followed by a batch mean. No loss is applied to historical image blocks.

Stage 2: prompt expansion. We adapt how the VLM turns a request into the visual instruction executed by the DiT. Both components are jointly adapted with a learning rate of 5×10^{-6}. AR loss covers only <prompt> content; tags, references, reasoning, user text, and delimiters receive no direct AR supervision. All tokens remain in context, and FM gradients reach the VLM through continuous features. System instructions specify call syntax using pretrained instruction following. They can request task-dependent formats, such as natural language for general edits or structured descriptions of coordinates and state changes for layout or board tasks. Each task's format can be specified separately, without changing other tasks' instruction templates.

For sample i, with mask m_{ij} and token loss ℓ_{ij}, the objectives are

$$\mathcal{L}_{\mathrm{AR}} = \frac{1}{B} \sum_{i=1}^{B} \frac{\sum_j m_{ij} \ell_{ij}}{\sum_j m_{ij}}, \qquad \mathcal{L}_{\mathrm{S2}} = \mathcal{L}_{\mathrm{AR}} + \mathcal{L}_{\mathrm{FM}}. \tag{1}$$

Both stages use a global batch size of 512. AR and FM have unit coefficients with no further magnitude matching. Their values stabilize around 1.5 and 0.15, respectively; FM constrains conditioning drift during prompt adaptation.

4 RESULTS

4.1 MAIN RESULTS ON PUBLIC BENCHMARKS

We evaluate UEval (Li et al., 2026), RISEBench (Zhao et al., 2025), GenEval 2 (Kamath et al., 2025), and GEdit-Bench (Liu et al., 2025). These benchmarks assess coordinated text–image output, reasoning-informed editing, prompt fulfillment, and general editing quality. Public UMMs are the main comparison group; specialist models and APIs provide capability references. Our GEdit evaluation uses the English full set (606 examples) and GPT-4.1. Higher scores are better.

Table 1: Unified image-text generation on UEval. **Bold**: best listed open UMM; underline: best listed model. Higher is better.

Model	Space	Textbook	Diagram	Paper	Art	Life	Tech	Exercise	Avg
Proprietary references									
GPT-5-Thinking	84.0	78.0	67.8	51.9	67.8	63.8	57.0	61.4	66.4
Gemini-2.5-Flash	78.0	74.0	66.4	71.6	66.6	63.0	58.2	50.0	66.0
Open unified multimodal models									
MMaDA	10.8	20.0	14.2	13.3	15.7	15.8	12.4	12.6	14.4
Show-o2	25.4	33.1	33.2	17.4	25.6	15.6	17.4	13.1	22.6
Janus-Pro	21.0	31.0	37.4	15.2	26.4	23.0	17.6	11.5	22.9
BAGEL	29.8	42.5	37.2	20.0	39.0	33.6	24.8	21.4	31.0
Emu3.5	**59.1**	57.4	41.1	31.6	**59.3**	**62.0**	37.0	**45.4**	49.1
Qwen-Image-Omni (ours)	54.3	**66.4**	**50.2**	**47.2**	56.1	55.7	**37.5**	36.9	**50.5**

Interleaved generation. UEval evaluates eight domains. Qwen-Image-Omni averages 50.5, exceeding the listed open UMMs, including Emu3.5 at 49.1. Its strongest relative results include Textbook, Diagram, Paper, and Tech, which require coordinated information-dense images and text.

Table 2: Reasoning-informed editing on RISEBench (%, higher is better). Bold: best listed UMM. †: paper-reported result; ‡: recomputed over 360 examples from reported category scores. §: source Table 19 values, retained verbatim; the reported overall (9.4) exceeds all category scores, an unresolved inconsistency in the source.

Model	Temporal	Causal	Spatial	Logical	Overall
Proprietary references					
GPT-Image-2	45.9	66.7	50.0	34.1	49.4
Gemini-2.5-Flash-Image	29.4	48.9	37.0	18.8	33.9
GPT-Image-1	36.5	34.4	37.0	10.6	30.0
Specialized editor					
Qwen-Image-Edit-2511	21.2	18.9	31.0	4.7	19.4
Native unified multimodal models					
BAGEL + WEAVE-100k‡	4.7	6.7	21.0	2.4	9.2
InternVL-U + CoT§	4.7	7.8	1.8	5.9	9.4
BAGEL + CoT	11.8	27.8	21.0	1.2	15.8
SenseNova-U1†	22.4	33.3	27.0	11.8	23.9
SenseNova-U1 + CoT†	31.8	33.3	27.0	**15.3**	26.9
Qwen-Image-Omni (ours)	**42.4**	**44.4**	**43.0**	7.1	**34.7**

Reasoning-informed editing. On RISEBench, Qwen-Image-Omni scores 34.7 overall versus 26.9 for SenseNova-U1 + CoT and 19.4 for Qwen-Image-Edit-2511. Logical editing remains challenging (7.1).

Table 3: Image generation and editing. GenEval 2 reports Soft-TIFA scores (0–100); GEdit-Bench-EN reports semantic consistency (G_{SC}), perceptual quality (G_{PQ}), and overall quality (G_{O}) (0–10). Higher is better; bold highlights our model. –: no verified result for the metric/setting.

Model	GenEval 2		GEdit-Bench-EN		
	AM	GM	G_{SC}	G_{PQ}	G_{O}
Proprietary reference					
Gemini-2.5-Flash-Image	82.8	44.6	7.41	7.96	7.10
Specialized generation / editing references					
CMO	80.0	34.0	–	–	–
Qwen-Image-2.0 (base)	79.7	41.0	–	–	–
Step1X-Edit-v1.1	–	–	7.66	7.35	6.97
Open unified multimodal models					
Janus	52.9	11.0	–	–	–
Show-o	63.4	16.0	–	–	–
HermesFlow	63.5	16.1	–	–	–
BAGEL	–	–	7.36	6.83	6.52
+ CoT	70.9	23.1	–	–	–
InternVL-U + CoT	–	–	–	–	6.88
Emu3.5	–	–	8.11	7.70	7.59
SenseNova-U1	–	–	8.27	7.49	7.47
Qwen-Image-Omni (ours)	**83.3**	**48.9**	**8.08**	**8.09**	**7.61**

Generation and general editing. GenEval 2 improves from the base model's 79.7/41.0 AM/GM to 83.3/48.9, gaining 3.6 and 7.9 points, respectively. The improvement extends to GM, which penalizes poorly satisfied atomic prompt conditions, supporting more complete fulfillment of detailed visual instructions. Table 3 reports an overall GEdit-Bench-EN score of 7.61 for Qwen-Image-Omni, compared with 7.59 for Emu3.5 and 7.47 for SenseNova-U1.

Internal reasoning in image content. Figure 4 illustrates how the VLM expands implied requirements. For continued photography in heavy rain, Qwen-Image-Omni includes protective equipment, correcting the base output's exposed cameras and lenses. The request specifies a practical goal. The VLM expands it into the equipment and visual details needed to depict that goal.

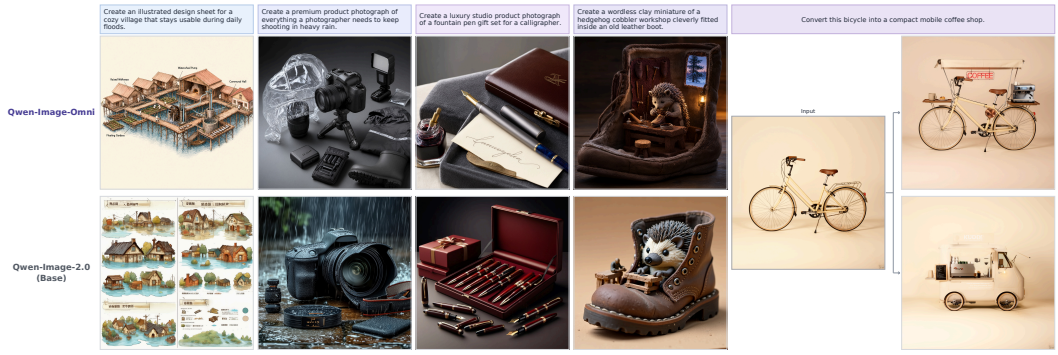

Figure 4: Qualitative effects of internal instruction expansion under identical user inputs, without an external prompt enhancer. Qwen-Image-Omni (top) includes rain protection in the second column, addressing the goal of continued photography. The base model is below. Four columns show generation; the last edits the same bicycle.

4.2 Ablation Studies

We use ComplexEval to examine the visual-call interface and supervision scope during multi-turn execution. Table 4 complements public-benchmark results with comparisons among variants of our model.

Tasks and interaction. The set contains 40 visual-design, 40 long-history editing, and 32 interleaved-generation cases, totaling 430 prescribed image steps. Cases were authored independently of training trajectories, with held-out reference assets; goals, budgets, and checklists were fixed before generation. Design and interleaving receive a system prompt, goal, and initial references once, then generate their own calls. Only historical editing receives a new instruction each turn. System instructions require valid skill calls, explicit reference roles and preservation constraints, and inspection of returned images. Design goals may specify an operation order, as in Figure 1. Figure 6 summarizes these inputs.

Scoring and human assessment. A step succeeds only if all its approximately four binary checks pass. For example, a bicycle edit checks bicycle presence, identity from R1, clothing from R2, and a bridge background. Reference correctness is judged from images; purely textual requirements use text. Historical steps comprise 53 generation, 68 single-reference, and 41 multi-reference steps, counting the editing target among references. Complete-task checks cover the requested final state and mandatory milestones for design, designated references and still-active final constraints for editing, and section completeness, text–image correspondence, and continuity for interleaving. Superseded editing constraints are excluded; prescribed round budgets are checked.

To support reliable visual judgments, two annotators independently score every trajectory with model identities and variant names hidden and presentation order randomized. A third reviews judgments and resolves disagreements against outputs, references, and task checklists. All reported scores use the final reviewed labels. Step rates pool prescribed steps, retaining failed and missing steps; overall rates weight cases equally and average four runs, rather than multiplying step rates. The independent task counts remain 40/40/32.

Settings and variants. All multi-turn variants use temperature 0.7, top-p 0.9, at most 4,096 text tokens per turn, and 40 Euler steps with CFG 4.0 at 1024^2. Four fixed seeds are paired by case; each run develops its own history and stops at its call budget, without retries or best-of-four selection. Coarse uses a different instruction format, described below.

Coarse uses a single `<general>` tag during training and inference, removes the structured `<ref>` block, and describes reference-image positions in natural language. Training data, both stages' training budgets, and evaluation call budgets match Full, providing a matched test of the combined skill-and-reference interface. Full-sequence AR starts from Full's Stage 1 checkpoint and also supervises reasoning, tags, and references; loss reduction, FM, and Stage 2 budget are unchanged. Stage 1 only omits the second stage and has fewer updates. Base uses unadapted weights and one call, scored only on terminal design checks.

Table 4: Ablations on ComplexEval by task family. Overall measures complete-task success; long-history step success is separated into generation (Gen.), single-reference editing (Single), and multi-reference editing (Multi). Scores are success rates in $[0, 1]$, averaged over four generations per case; higher is better. Bold marks the best score in each column. Base is a final-image-only reference with unadapted weights; dashes indicate unevaluated settings.

Configuration	Design (40)	Long-history editing (40)				Interleaved (32)
	Overall	Step success			Overall	Overall
		Gen.	Single	Multi		
Full model	**0.70**	**0.87**	**0.83**	**0.68**	**0.41**	0.59
Coarse interface	0.62	0.85	0.71	0.59	0.29	**0.62**
Stage 1 only	0.66	0.82	0.77	0.63	0.39	0.52
Full-sequence AR	0.43	0.85	0.74	0.62	0.33	0.55
Base (single-round)	0.46	–	–	–	–	–

Structured calls versus general instructions. Under matched budgets, Full raises design success from Coarse's 0.62 to 0.70 and historical-editing success from 0.29 to 0.41. Gains are larger for single-reference (0.71 to 0.83) and multi-reference editing (0.59 to 0.68) than generation (0.85 to 0.87). These results establish the practical value of explicit operation types and image references for the tested visual programs. In design, each transition between text insertion, object editing, and restyling must preserve earlier decisions; the interface specifies both the intended change and its visual sources. Coarse remains stronger on interleaving (0.62 versus 0.59), indicating task-dependent benefits.

Prompt supervision. Prompt-only AR reaches 0.70 design success versus 0.43 for full-sequence AR under matched Stage 2 budgets, supporting direct text supervision focused on visual instructions. Stage 2 improves design success from 0.66 to 0.70 and historical-editing success from 0.39 to 0.41. Figure 5 documents FM convergence over eleven target rounds.

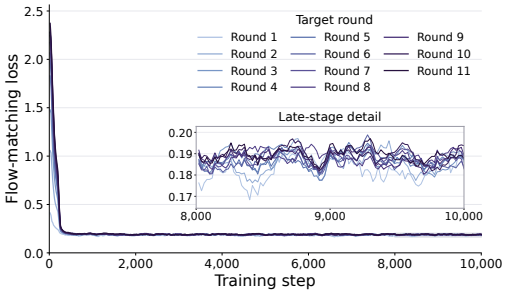

Figure 5: Stage 1 FM loss across target rounds 1–11 over the displayed training interval, with an inset showing steps 8,000–10,000.

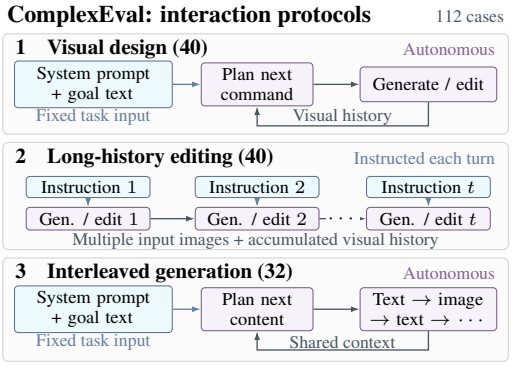

Figure 6: Interaction protocols for the ablation tasks. Blue: external inputs; purple: model execution. Design and interleaving generate their own calls from a fixed system prompt and goal. Historical editing receives per-turn instructions and binds references while preserving the current state.

Action and completion. The goal-driven system prompt distinguishes an image action from a completion decision. A generation or editing block requests image synthesis; the model must inspect the returned image before deciding what to do next. A `<done>` block identifies the actual images to deliver through their global IDs. A `<blocked>` block states unmet requirements when execution cannot continue, for example because an indispensable input is missing or the call budget is exhausted. Neither terminal block contains a visual prompt or requests image synthesis. References must point to images already available in history, rather than anticipated outputs. These instructions define the interaction contract; emitting a completion token does not itself establish visual task success.

5 Discussion

Qwen-Image-Omni combines shared visual history, structured natural-language calls, and two-stage adaptation to support complex visual programs. The same VLM constructs instructions and conditions their visual execution. Each call identifies the operation, its visual sources, and the content to preserve. The ability to inspect and reuse intermediate images opens several directions for Qwen-Image-Omni. A next step is to test whether the VLM can detect a failed edit and revise subsequent calls to recover. Longer design workflows could require revisiting earlier alternatives and combining references whose roles change across turns. Expanding the skill repertoire and using system prompts to define task-specific interfaces could also let users intervene in a visual program, revise its goal, and continue from an existing result.

AI USE STATEMENT

In this work, we used generative AI tools to assist with research code implementation and the analysis of experimental results. We did not use generative AI tools to develop conceptual frameworks, formulate hypotheses, or design experiments; tasks related to theoretical proofs are not applicable to this work. Additionally, we used generative AI tools to polish the manuscript for clarity, readability, and linguistic consistency. We manually inspected AI-assisted code for correctness, verified AI-assisted result analyses against experimental records and reported calculations, and reviewed all text revisions for accuracy and consistency with our intended scientific meaning. We take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the assistance of generative AI.

ETHICS STATEMENT

Qwen-Image-Omni supports image generation, editing, and the reuse of historical images within visual workflows. These capabilities can support creative and educational applications, but may also facilitate misleading image manipulation. In particular, preserving a person's identity while changing their appearance or surroundings could enable impersonation or fabricated depictions. Responsible use requires respecting privacy, consent, and intellectual property associated with training data and reference images. Generated or substantially edited images should be identified as such when they could otherwise be mistaken for authentic records. The capabilities evaluated in this work concern visual task execution and should not be interpreted as evidence of safety against these forms of misuse.

REPRODUCIBILITY STATEMENT

The main paper describes the shared-history architecture, structured skill calls, training-data construction, two-stage optimization, and the settings used for the ablation studies. For ComplexEval, it specifies the three interaction protocols, step-level and complete-task success criteria, aggregation over four generations per case, and the human evaluation procedure. The reported inference settings include a classifier-free guidance scale of 4.0, 40 sampling steps, a text-generation temperature of 0.7, and a maximum of 4,096 text tokens per turn. Exact reproduction additionally requires access to the pretrained checkpoints, training data, and complete evaluation assets, including task instructions, reference images, and case-specific checklists. These dependencies should be considered when reproducing the training pipeline or comparing results under the ComplexEval protocol.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Feifei Bian, Zhimin Zheng, Wei Deng, Daiguo Zhou, and Jian Luan. RS-Gen: A multi-stage agentic framework for reasoning and search-augmented image generation. *arXiv preprint arXiv:2606.23221*, 2026.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. FLUX.1 Kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. InstructPix2Pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18392–18402, 2023.

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

Wei Chen, Lin Li, Yongqi Yang, Bin Wen, Fan Yang, Tingting Gao, Yu Wu, and Long Chen. CoMM: A coherent interleaved image-text dataset for multimodal understanding and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8073–8082, 2025a.

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*, 2025b.

Wei Chow, Jiachun Pan, Yongyuan Liang, Mingze Zhou, Xue Song, Liyu Jia, Saining Zhang, Siliang Tang, Juncheng Li, Fengda Zhang, Weijia Wu, Hanwang Zhang, and Tat-Seng Chua. WEAVE: Unleashing and benchmarking the in-context interleaved comprehension and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15343–15353, 2026.

Yufeng Cui, Honghao Chen, Haoge Deng, Xu Huang, Xinghang Li, Jirong Liu, Yang Liu, Zhuoyan Luo, Jinsheng Wang, Wenxuan Wang, Yueze Wang, Chengyuan Wang, Fan Zhang, Yingli Zhao, Ting Pan, Xianduo Li, Zecheng Hao, Wenxuan Ma, Zhuo Chen, Yulong Ao, Tiejun Huang, Zhongyuan Wang, and Xinlong Wang. Emu3.5: Native multimodal models are world learners. *arXiv preprint arXiv:2510.26583*, 2025.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Haiwen Diao, Jiahao Wang, Chenjing Ding, Hanming Deng, Jiangnan Chen, Ruixi Zhang, Ruohui Wang, Wenwen Tong, Xiangyu Fan, Yubo Wang, et al. SenseNova-U1.5: Towards native unified visual intelligence. *arXiv preprint arXiv:2609.11929*, 2026a.

Haiwen Diao, Penghao Wu, Hanming Deng, Jiahao Wang, Shihao Bai, Silei Wu, Weichen Fan, Wenjie Ye, Wenwen Tong, Xiangyu Fan, et al. SenseNova-U1: Unifying multimodal understanding and generation with NEO-unify architecture. *arXiv preprint arXiv:2605.12500*, 2026b.

Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, Xiangwen Kong, Xiangyu Zhang, Kaisheng Ma, and Li Yi. DreamLLM: Synergistic multimodal comprehension and creation. In *International Conference on Learning Representations*, pp. 6666–6702, 2024.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12606–12633, 2024.

Tsu-Jui Fu, Wenze Hu, Xianzhi Du, William Yang Wang, Yinfei Yang, and Zhe Gan. Guiding instruction-based image editing via multimodal large language models. In *International Conference on Learning Representations*, 2024.

Yuying Ge, Sijie Zhao, Ziyun Zeng, Yixiao Ge, Chen Li, Xintao Wang, and Ying Shan. Making LLaMA SEE and draw with SEED tokenizer. In *International Conference on Learning Representations*, pp. 35206–35231, 2024a.

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*, 2024b.

Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14953–14962, 2023.

Yuzhou Huang, Liangbin Xie, Xintao Wang, Ziyang Yuan, Xiaodong Cun, Yixiao Ge, Jiantao Zhou, Chao Dong, Rui Huang, Ruimao Zhang, and Ying Shan. SmartEdit: Exploring complex instruction-based image editing with multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8362–8371, 2024.

Amita Kamath, Kai-Wei Chang, Ranjay Krishna, Luke Zettlemoyer, Yushi Hu, and Marjan Ghazvininejad. GenEval 2: Addressing benchmark drift in text-to-image evaluation. *arXiv preprint arXiv:2512.16853*, 2025.

Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014.

Jing Yu Koh, Daniel Fried, and Ruslan Salakhutdinov. Generating images with multimodal language models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 21487–21506, 2023.

Siqi Kou, Jiachun Jin, Zhihong Liu, Chang Liu, Ye Ma, Jian Jia, Quan Chen, Peng Jiang, and Zhijie Deng. Orthus: Autoregressive interleaved image-text generation with modality-specific heads. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 31619–31633. PMLR, 2025.

Bo Li, Yida Yin, Wenhao Chai, Xingyu Fu, and Zhuang Liu. UEval: A benchmark for unified multimodal generation. In *European Conference on Computer Vision*, 2026.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, Guopeng Li, Yuang Peng, Quan Sun, Jingwei Wu, Yan Cai, Zheng Ge, Ranchen Ming, Lei Xia, Xianfang Zeng, Yibo Zhu, Binxing Jiao, Xiangyu Zhang, Gang Yu, and Daxin Jiang. Step1X-Edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023.

Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai Yu, Liang Zhao, Yisong Wang, Jiaying Liu, and Chong Ruan. JanusFlow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7739–7751, 2025.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. HuggingGPT: Solving AI tasks with ChatGPT and its friends in Hugging Face. In *Advances in Neural Information Processing Systems*, volume 36, pp. 38154–38180, 2023.

Min Shi, Xiaohui Zeng, Jiannan Huang, Yin Cui, Francesco Ferroni, Jialuo Li, Zhaoshuo Li, Yogesh Balaji, Haoxiang Wang, Tsung-Yi Lin, Xiao Fu, Yue Zhao, Chieh-Yun Chen, Ming-Yu Liu, and Humphrey Shi. DuoGen: Towards autonomous interleaved multimodal generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21901–21911, 2026.

Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14398–14409, 2024a.

Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Emu: Generative pretraining in multimodality. In *International Conference on Learning Representations*, pp. 12352–12380, 2024b.

Dídac Surís, Sachit Menon, and Carl Vondrick. ViperGPT: Visual inference via Python execution for reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11888–11898, 2023.

Changyao Tian, Xizhou Zhu, Yuwen Xiong, Weiyun Wang, Zhe Chen, Wenhai Wang, Yuntao Chen, Lewei Lu, Tong Lu, Jie Zhou, Hongsheng Li, Yu Qiao, and Jifeng Dai. MM-Interleaved: Interleaved image-text generative modeling via multi-modal feature synchronizer. *arXiv preprint arXiv:2401.10208*, 2024.

Changyao Tian, Danni Yang, Guanzhou Chen, Erfei Cui, Zhaokai Wang, Yuchen Duan, Penghao Yin, Sitao Chen, Ganlin Yang, Mingxin Liu, et al. InternVL-U: Democratizing unified multimodal models for understanding, reasoning, generation and editing. *arXiv preprint arXiv:2603.09877*, 2026.

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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17001–17012, 2025.

Chonghuinan Wang, Zhikai Chen, Chunwei Wang, Yecong Wan, Junwei Yang, Zhixin Wang, Wei Zhang, Jiaqi Xu, Renjing Pei, Xiaohe Wu, Fan Li, and Wangmeng Zuo. Illuminating unified multimodal model for free-form interleaved text-image generation. In *European Conference on Computer Vision*, 2026a.

Dianyi Wang, Chaofan Ma, Feng Han, Size Wu, Wei Song, Yibin Wang, Zhixiong Zhang, Tianhang Wang, Siyuan Wang, Zhongyu Wei, and Jiaqi Wang. UniReason 1.0: A unified reasoning framework for world knowledge aligned image generation and editing. *arXiv preprint arXiv:2602.02437*, 2026b.

Xinlong Wang, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Zhen Li, Yuqi Wang, Qiying Yu, Yingli Zhao, Yulong Ao, Xuebin Min, Chunlei Men, Boya Wu, Bo Zhao, Bowen Zhang, Liangdong Wang, Guang Liu, Zheqi He, Xi Yang, Jingjing Liu, Yonghua Lin, Zhongyuan Wang, and Tiejun Huang. Multimodal learning with next-token prediction for large multimodal models. *Nature*, 650(8101):327–333, 2026c.

Jungmyung Wi, Hyunsoo Kim, and Donghyun Kim. Correlation-weighted multi-reward optimization for compositional generation. In *European Conference on Computer Vision*, pp. 93–111, 2026.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-Image technical report. *arXiv preprint arXiv:2508.02324*, 2025a.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, and Ping Luo. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12966–12977, 2025b.

Chenyuan Wu, Jiahao Wang, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, Ziyi Xia, Ze Liu, Chaofan Li, Haoge Deng, Kun Luo, Bo Zhang, Jiajun Zhang, Dong Liu, Defu Lian, Xinlong Wang, Zhongyuan Wang, Tiejun Huang, and Zheng Liu. OmniGen2: Towards instruction-aligned multimodal generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21964–21975, 2026.

Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, Song Han, and Yao Lu. VILA-U: a unified foundation model integrating visual understanding and generation. In *International Conference on Learning Representations*, pp. 93620–93638, 2025c.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13294–13304, 2025.

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. In *International Conference on Learning Representations*, pp. 28240–28264, 2025a.

Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 47490–47518, 2025b.

Jinbo Xing, Zeyinzi Jiang, Yuxiang Tuo, Chaojie Mao, Xiaotang Gai, Xi Chen, Jingfeng Zhang, Yulin Pan, Zhen Han, Jie Xiao, Keyu Yan, Chenwei Xie, Chongyang Zhong, Kai Zhu, Tong Shen, Lianghua Huang, Yu Liu, and Yujiu Yang. Wan-Weaver: Interleaved multi-modal generation via decoupled training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 36267–36278, 2026.

Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. MMaDA: Multimodal large diffusion language models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 138867–138907, 2025a.

Ling Yang, Xinchen Zhang, Ye Tian, Shiyi Zhang, Chenming Shang, Minghao Xu, Wentao Zhang, and Bin Cui. HermesFlow: Seamlessly closing the gap in multimodal understanding and generation. In *Advances in Neural Information Processing Systems*, volume 38, pp. 62248–62272, 2025b.

Yuxiao Ye, Haoran He, Fangyuan Kong, Xintao Wang, Pengfei Wan, Kun Gai, and Ling Pan. Edit-R2: Context-aware reinforcement learning for multi-turn image editing. *arXiv preprint arXiv:2606.05950*, 2026.

Shengming Yin, Zekai Zhang, Zecheng Tang, Kaiyuan Gao, Xiao Xu, Kun Yan, Jiahao Li, Yilei Chen, Yuxiang Chen, Heung-Yeung Shum, Lionel M. Ni, Junyang Lin, and Chenfei Wu. Qwen-Image-Layered: Towards inherent editability via layer decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16196–16205, 2026a.

Zijin Yin, Tiankai Hang, Yiji Cheng, Shiyi Zhang, Runze He, Yu Xu, Chunyu Wang, Bing Li, Zheng Chang, Kongming Liang, Qinglin Lu, and Zhanyu Ma. Generative visual chain-of-thought for image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 4657–4667, 2026b.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26125–26135, 2025.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023.

Bing Zhao, Chenfei Wu, Deqing Li, Hao Meng, Jiahao Li, Jie Zhang, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kuan Cao, Kun Yan, Liang Peng, Lihan Jiang, Niantong Li, Ningyuan Tang, Shengming Yin, Tianhe Wu, Xiao Xu, Xiaoyue Chen, Xihua Wang, Yan Shu, Yanran Zhang, Yi Wang, Yilei Chen, Ying Ba, Yixian Xu, Yujia Wu, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhendong Wang, Zihao Liu, Zikai Zhou, An Yang, Chen Cheng, Chenxu Lv, Dayiheng Liu, Fan Zhou, Hantian Xiong, Hongzhu Shi, Hu Wei, Huihong Zhao, Ivy Liu, Jianwei Zhang, Jiawei Zhang,

Kai Chen, Kang He, Levon Xue, Lin Qu, Linhan Tang, Luwen Feng, Minggang Wu, Minmin Sun, Na Ni, Rui Men, Shuai Bai, Sishou Zheng, Tao Lan, Tianqi Zhang, Tingkun Wen, Wei Wang, Weixu Qiao, Weiyi Lu, Wenmeng Zhou, Xiaodong Deng, Xiaoxiao Xu, Xinlei Fang, Xionghui Chen, Yanan Wang, Yang Fan, Yichang Zhang, Yixuan Xu, Yu Wu, Zhiyuan Ma, and Zhizhi Cai. Qwen-Image-2.0 technical report. *arXiv preprint arXiv:2605.10730*, 2026.

Xiangyu Zhao, Peiyuan Zhang, Kexian Tang, Xiaorong Zhu, Hao Li, Wenhao Chai, Zicheng Zhang, Renqiu Xia, Guangtao Zhai, Junchi Yan, Hua Yang, Xue Yang, and Haodong Duan. Envisioning beyond the pixels: Benchmarking reasoning-informed visual editing. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

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. In *International Conference on Learning Representations*, pp. 6446–6469, 2025.