

Figure 1: **Gallery for V-protocol.**

CAN IMAGE AGENTS COMPLETE VISUAL COMMISSIONS? A BENCHMARK AND MODULAR FRAMEWORK

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual commissions require image agents to integrate heterogeneous source materials, preserve requirements across dependent stages, and produce coordinated deliverables. Existing evaluations examine specific capabilities and execution settings, leaving end-to-end commission completion less understood. We introduce Visual Commission Bench (VCB), comprising 600 tasks across 12 categories and three complementary tiers, from capability-focused tasks to complete commissions. VCB uses 5,039 requirement checks and separate visual quality ratings to assess content correctness, visual realization, constraint preservation, and delivery completeness without prescribing an internal workflow. We present V-Protocol, a reference agent that dynamically composes reusable task, cognitive, and expression modules at inference time. Task modules organize delivery, cognitive modules derive content, and expression modules guide visual realization. The modules exchange structured intermediate artifacts, while a shared execution harness tracks artifact revisions and dependencies, manages tool execution, and checks completion. With the same controller and image generator as the agent baselines, V-Protocol achieves 69.0% strict task success on VCB, compared with 42.0% for SCOPE. On Mind-Bench, strict accuracy increases from 39.2% to 45.4%. The gallery for V-protocol could be found in Figure 1.

1 INTRODUCTION

Modern image models follow detailed instructions Wu et al. (2025) to produce high-quality visuals, with growing use in visual production tasks. A visual commission, however, may begin with incomplete requirements and heterogeneous source materials and involve multiple dependent stages before final delivery. For example, preparing promotional materials for a tea house may require identifying valid content, deciding how to present it across different assets, preserving relevant con-

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.

straints through generation and revision, and producing a complete and coherent set of deliverables. Much of this process is still coordinated manually, while text-to-image generation serves as only one part of the workflow. We study whether image agents can carry out such visual commissions, from interpreting source materials through to producing the required deliverables, within a given execution budget. Figure 2 illustrates representative tasks and deliverables.

Existing evaluations have extended image generation beyond direct prompt following by introducing tasks that require reasoning, search, or planning Zhang et al. (2026); Ren et al. (2026); Tang et al. (2026); Ki et al. (2025). Despite these broader demands, existing evaluations typically focus on particular capabilities or execution settings within the broader visual commission process. Such settings allow specific capabilities and execution demands to be examined under controlled conditions. It therefore remains unclear whether success in these bounded settings carries over to end-to-end visual commission completion.

Existing evaluations have extended image generation beyond direct prompt following by introducing tasks that require require reasoning, search, or planning Zhang et al. (2026); Ren et al. (2026); Tang et al. (2026); Ki et al. (2025). These settings provide valuable tests of capabilities and execution demands involved in visual production. A further question is whether success in these settings translates into reliable completion of visual commissions, where agents must integrate requirements and evidence distributed across source materials, carry intermediate decisions and constraints through dependent subtasks, and coordinate multiple deliverables. Evaluating these demands together allows us to examine how constituent capabilities support completion of the overall commission.

To address this question, we introduce **Visual Commission Bench (VCB)**, comprising 600 tasks across 12 categories and three complementary tiers: Specific-Capability (S), Multi-Stage (M), and Commission-Level (C). S tasks isolate capabilities relevant to visual task completion, such as reasoning, planning, information extraction, and retrieval. M tasks examine how intermediate results, decisions, and constraints are carried across dependent stages. C tasks evaluate complete visual commissions from source materials to coordinated delivery. The three tiers distinguish evaluation scope and task dependencies rather than intrinsic difficulty or tool-call counts, allowing us to study both individual capabilities and how they contribute to end-to-end task completion. All tasks are evaluated through their delivered artifacts without prescribing an internal workflow. 5,039 carefully designed requirement checks are used to measure requirement-level fulfillment and derive task-level success.

VCB further reveals execution challenges that become salient at the commission level. Some trajectories make little observable task progress despite substantial interaction, with agents repeatedly revisiting guidance and source materials, checking unchanged states, or recovering from similar interface errors. These behaviors suggest that commission-level execution depends not only on the capabilities available to the agent, but also on how appropriate task procedures are selected, reused, and coordinated throughout execution Li et al. (2026c); Liu et al. (2026); Dong et al. (2026); Tan et al. (2026). This motivates us to examine structured support for both task guidance and execution management. Preliminary runs on VCB reveal execution difficulties in carrying visual commissions through to completion.

We present **V-Protocol** as a reference agent for end-to-end visual commission completion. V-Protocol organizes reusable task knowledge into task, cognitive, and expression modules, and dynamically selects compatible combinations based on effect profiles estimated on separate development tasks. Task modules specify deliverables and execution dependencies, cognitive modules derive content from the inputs, and expression modules produce layout and rendering specifications. Selected modules exchange structured intermediate artifacts, while a shared harness Pan et al. (2026); Dantanarayana et al. (2026); Lu et al. (2026); Xia et al. (2026) manages stage transitions, tool execution, and completion checks. Together, these components support controlled study of visual commission completion while keeping the underlying foundation models fixed.

2 Related Work

Benchmarks for image agents. WISE evaluates world knowledge integration Niu et al. (2026), while Mind-Bench and KnowGen assess reasoning and search in image generation He et al. (2026a); Feng et al. (2026). Gen-Arena measures entity and constraint satisfaction Ren et al. (2026), and

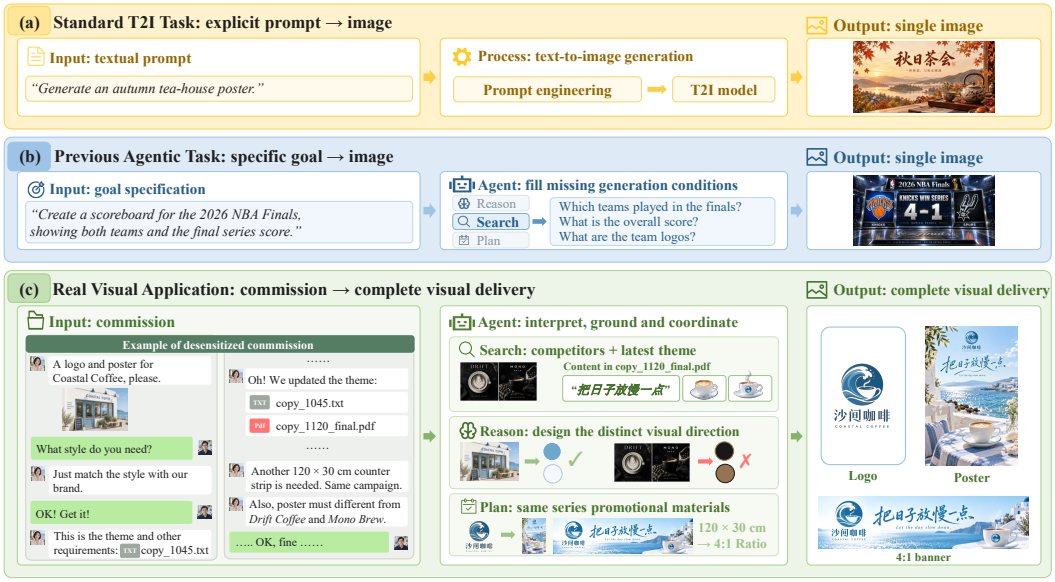

Figure 2: **From image generation to visual commission completion.** (a). An explicit prompt guides image synthesis. (b). A specified generation goal is supplemented with missing information through reasoning, search, and planning. (c). A commission and its source materials require the agent to determine content and design, preserve active constraints, and coordinate the requested deliverables. VCB evaluates content correctness, visual realization, constraint preservation, and delivery completeness. The panels illustrate task settings rather than mutually exclusive method categories.

IA-Bench covers planning, reasoning, search, and memory Zhang et al. (2026). GraphicBench and IGenBench examine graphic design and infographic reliability, respectively Ki et al. (2025); Tang et al. (2026). VCB evaluates complete visual commissions that require agents to integrate distributed source materials, preserve requirements across dependent stages, and coordinate multiple deliverables. It assesses the delivered artifacts without prescribing an internal workflow.

Agentic image generation. Gen-Searcher and Mind-Brush integrate retrieval and reasoning into generation Feng et al. (2026); He et al. (2026a). GenEvolve learns tool use from visual experience Chen et al. (2026b), and GEMS integrates memory and on-demand skills He et al. (2026b). SCOPE maintains evolving semantic specifications for conditional skill orchestration Ren et al. (2026), while Qwen-Image-Agent constructs generation context through planning and grounding Zhang et al. (2026). Other approaches use executable sketches Ye et al. (2026), adaptive workflow routing Chen et al. (2026a), or evolving skill harnesses Li et al. (2026d). V-Protocol coordinates reusable task, cognitive, and expression modules through structured intermediate artifacts, linking content derived from source materials to its use across visual deliverables.

Skills and execution management. AgentSkillOS and SkillSelect-Serve address skill organization and budget-aware selection Li et al. (2026a); Zheng et al. (2026); SkillSmith models complementarity and conflict among skills Wei et al. (2026). GraSP composes skills into typed dependency graphs with verification and local repair Xia et al. (2026). Natural-language harnesses expose reusable execution policies Pan et al. (2026), while SIGIL and ContractSkill make procedural guidance executable and checkable Dantanarayana et al. (2026); Lu et al. (2026). V-Protocol builds on these mechanisms to track artifact dependencies and completion conditions, leaving content and design decisions to the model. Visual checks assess requirement satisfaction alongside the harness's checks on execution state.

Table 1: **Comparison of benchmark scale and evaluation focus.** Checks denote task-specific verification items, while rating dimensions denote graded evaluation criteria. VCB broadens the evaluation toward end-to-end visual commission completion.

Benchmark	# Tasks	Checks / Ratings	Primary focus
WISE	1,000	3-dim ratings	World-knowledge reasoning
Mind-Bench	500	1,633 checks	Reasoning and search
KnowGen	630	4-dim ratings	Search-grounded generation
Gen-Arena	300	2,533 checks	Entity and constraint satisfaction
IA-Bench[1]	730	1,801 checks	Planning, reasoning, search, and memory
VCB (Ours)	**600**	**5,039 checks + 4-dim ratings**	**Distributed-source integration, dependent subtask execution, and coordinated multi-artifact delivery**

[1] IA-Bench is a closed source internal evaluation benchmark.

3 Visual Commission Benchmark

Visual Commission Benchmark (VCB) evaluates the capabilities required for visual commission completion at different levels. Tasks require agents to identify valid content from source materials, decide how to present it across different assets, preserve relevant constraints through generation and revision, and produce a complete and coherent set of deliverables. Evaluation is based on the submitted artifacts, covering content correctness, visual realization, constraint preservation, and delivery completeness without prescribing an internal workflow. Table 1 summarizes how VCB relates to existing evaluation settings. Figure 5 in the appendix provides an overview of V-Protocol, including adaptive module composition, protocol execution, and artifact revision management.

3.1 Task Scope and Public Inputs

As shown in Figure 3, VCB comprises 600 tasks across 12 categories, with 50 tasks per category, organized into three complementary tiers: Specific-Capability (S), Multi-Stage (M), and Commission-Level (C). S tasks emphasize a focal capability: reasoning, planning, long-context information extraction, or retrieval. M tasks examine how capabilities work together through iterative editing, visual goal interpretation, diagnosis and repair, and evidence synthesis. C tasks involve visual argumentation, example-driven transfer, scientific visualization, and workspace-based design, requiring agents to make content and presentation decisions and organize the work needed for a complete delivery. The tiers provide a progression in evaluation scope and capability coordination without prescribing a fixed workflow or number of model and tool calls.

Each task provides a user request, access to its source materials, and designated output locations. Materials may include documents, tables, reference images, demonstrations, interaction histories, and workspace files. Task-relevant information and constraints may be distributed across the source materials, requiring agents to identify and integrate them. A task may require a single image, a sequence of edited images, or a coordinated set of visual artifacts. Private evaluation annotations specify how to assess requirements grounded in the request and materials, without adding undisclosed task goals or prescribing a unique valid solution. All tasks are self-contained evaluation packages and do not require additional user clarification during execution.

3.2 Construction and Annotation

Task sources and diversity. VCB draws on hard cases from existing text-to-image Li et al. (2026b) and image-agent benchmarks Niu et al. (2026); He et al. (2026a); Feng et al. (2026), newly constructed tasks, everyday user requests, and completed design commissions. Real-world cases are anonymized and adapted into evaluation tasks while preserving their information dependencies, requirement changes, and delivery constraints. Diversity spans both content domains and task structures, including information selection, dependent operations, constraint preservation, and multi-artifact coordination. The resulting benchmark contains 600 tasks across 12 taskcategories and 45 subcategories, with 50 tasks per category and an equal split of 300 Chinese and 300 English tasks.



Figure 3: **Overview of Visual Commission Benchmark (VCB).** VCB contains 600 tasks across three tiers, 12 categories, and 45 subcategories. The S tier focuses on reasoning, planning, long-context extraction, and retrieval. The M tier combines capabilities through iterative editing, visual goal interpretation, diagnosis and repair, and evidence synthesis. The C tier evaluates integrated visual commissions through visual argumentation, example transfer, research visualization, and workspace design. Each panel presents an example input and output, with darker labels indicating the represented subcategories. Documents and workspaces are shown schematically, and the displayed outputs are illustrative rather than experimental results.

Task construction. Source cases are organized into evaluation tasks comprising a user request, source materials, and delivery requirements. Relevant information may be distributed across documents, images, demonstrations, and interaction histories, requiring agents to identify and integrate it. In total, VCB provides 4,288 source files containing approximately 9.81 million tokens of machine-readable text across task-specific materials and shared corpora. Token counts use `cl100k_base` and exclude text embedded in images. Existing design outcomes can inform task construction and review. For tasks that permit alternative designs, they serve as references rather than the only acceptable solutions.

Checklist construction and review. For each task, multiple models independently construct candidate checklists from its request and source materials. The participating models include GPT5.6-Sol, GPT6-Astra, GLM-5.3-Flash, and DeepSeek-v4.1-Flash. The candidate checklists are consolidated by merging overlapping checks, retaining complementary requirements, and resolving conflicting criteria. Each resulting check expresses a concrete judgment question and specifies the supporting evidence and images needed for evaluation. The models then assess the consolidated checks individually for correctness, grounding, and clarity, and examine the checklists for unsupported requirements, redundancy, and omissions. Human reviewers examine the proposed checks and model judgments to identify necessary revisions. The checklists are refined through repeated rounds of model assessment, human review, and revision.

3.3 External Evaluation

ReqPass and TaskPass. Evaluation is based on the artifacts saved at the designated output locations. Let $\mathcal{Q}_n$ denote the applicable checklist items for task n, and let $r_{nj} \in \{0, 1\}$ indicate whether item j passes. Let $v_n \in \{0, 1\}$ denote the task-level delivery check, which verifies that all required artifacts are saved as specified and satisfy the applicable file-level requirements. We report

requirement pass rate and strict task pass rate:

$$\begin{aligned} \text{ReqPass} &= 100\,\frac{\sum_{n=1}^{N}\left(v_n + \sum_{j\in\mathcal{Q}_n} r_{nj}\right)}{N + \sum_{n=1}^{N}|\mathcal{Q}_n|}, \\ \text{TaskPass} &= \frac{100}{N}\sum_{n=1}^{N} v_n \prod_{j\in\mathcal{Q}_n} r_{nj}. \end{aligned} \tag{1}$$

ReqPass pools applicable checklist items and one delivery check per task, while TaskPass requires all of them to pass. Checks that depend on a missing or invalid artifact fail; checks on other available artifacts remain eligible for partial credit. Conditional checks are included only when applicable, and revision checks follow the requirements active at the corresponding checkpoint.

Visual quality. Visual quality (VQ) assesses readability, layout, visual coherence, and aesthetic appeal separately from factual correctness. Each assessable output image receives a rating from 1 to 5 on each dimension. Let $\mathcal{O}_n$ be the set of required output images for task n, and let s_{nkd} be the rating of image k on dimension d. We report

$$\text{VQ} = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{|\mathcal{O}_n|}\sum_{k\in\mathcal{O}_n}\frac{1}{4}\sum_{d=1}^{4} s_{nkd}. \tag{2}$$

Missing or undecodable required images receive zero on all four dimensions. Ratings are averaged within each task before being averaged across tasks, giving each task equal weight.

Judgment protocol. File-level properties, such as existence, decodability, and specified dimensions, are checked programmatically. Qwen3.8-Flash-Next serves as the evaluator for both requirement satisfaction and visual quality, with separate inputs and criteria. For requirement evaluation, it receives the checklist question, its supporting criteria, and the images needed for that judgment. Source facts are incorporated into the criteria, while reference images are supplied when visual comparison is necessary. For VQ evaluation, it receives only the output image and the fixed quality rubric. Both evaluations are blinded to the identity of the evaluated system and exclude its execution trace. Evaluator failures or unresolved judgments are recorded separately and retried rather than treated as task failures or assigned zero VQ. Final scores are reported once all required judgments have been resolved.

4 V-PROTOCOL

V-Protocol composes reusable modules into executable programs for visual commissions. The model selects procedures and determines content and design; a shared harness manages artifact dependencies, tool execution, and completion checks.

4.1 MODULE REPRESENTATION

Task modules specify commission procedures and optional program templates; cognitive modules guide material interpretation, retrieval, reasoning, and synthesis; expression modules provide layout, style, and rendering guidance. Each module declares its family, capabilities, input and output types, applicability conditions, and procedural guidance. Registry metadata supports selection, with full guidance loaded on activation. Declared interfaces identify possible handoffs, while the model assesses their suitability for the current task.

4.2 ADAPTIVE MODULE COMPOSITION

The agent derives a task specification from the public request and inspected materials, recording the goal, inputs, active requirements, and deliverables. A suitable task template initializes the program; otherwise, the model constructs it directly. At step t, the program is

$$\mathcal{P}_t = (G_t, \mu_t), \qquad G_t = (\mathcal{V}_t, \mathcal{E}_t), \tag{3}$$

where nodes represent operations, edges record declared artifact dependencies, and μ_t records module assignments. Each operation specifies its objective, input and output artifacts, capability needs, and permitted actions. Source materials can supply inputs directly. Operations may use several modules, and modules may be reused across operations.

For each operation, the registry retrieves modules whose capabilities overlap its needs and orders candidates by configured priority and capability coverage. The model selects applicable guidance for the current objective and inputs, revising the program as inspection and generation reveal additional needs. Simple requests can use a minimal program. Module-effect estimates are used for offline diagnostics rather than runtime selection.

4.3 Protocol Execution

The harness validates the operation graph and artifact references. Before each action, it checks input availability, dependency freshness, permissions, and remaining budget. The model supplies content and tool arguments; the harness mediates calls and registers outputs. Node completion requires recorded artifacts and valid tool results. Program updates retain execution history and valid completed outputs while allowing future operations to change. Retries and repairs consume the same task budget.

Artifacts carry identifiers, revisions, and declared source dependencies. Dependency snapshots record the source versions and content hashes used by downstream operations. A change to a declared source makes the corresponding snapshot stale, preventing affected artifacts from being reused as current evidence until the dependency is resolved. Image edits record their source revision, and replacing an image at the same path creates a new revision.

Online verification compares the current image with active requirements and content checkpoints. Reports bind to the image identity, revision, content hash, and verification criteria; changes to the image or relevant criteria invalidate prior reports. The model uses feedback to choose edits, regeneration, or program revisions. Before delivery, the harness checks that required artifacts exist at their designated locations and satisfy runtime completion conditions. These checks concern recorded requirements and dependencies; semantic judgments remain the model's responsibility. Private benchmark annotations are unavailable during planning, verification, and repair, and final artifacts are assessed through the external evaluation protocol.

5 Experiments

5.1 Experimental Setup

We evaluate V-Protocol on VCB, WISE-Verified, and Mind-Bench. We follow each benchmark's evaluation protocol. For VCB, we report ReqPass, TaskPass, and visual quality (VQ), both overall and across the S, M, and C tiers. WISE-Verified uses binary WiScore, and Mind-Bench uses checklist-based strict accuracy (CSA).

We compare against direct Qwen-Image2.1 generation, Mind-Brush, GEMS, and SCOPE. Agent-based methods share a Qwen3.8-27B controller with medium thinking and a Qwen-Image2.1 image backend. Where web search is permitted, they use DuckDuckGo. Each method retains its own planning, generation, and revision procedure. Qwen3.8-Flash-Next evaluates VCB requirement satisfaction and visual quality using separate prompts. Published reference results are presented separately from experiments with shared backends. Additional adaptation, evaluation, and recovery details are provided in Appendix B.

5.2 Results and Analysis

VCB. Table 2 shows that V-Protocol achieves the highest overall scores, with 94.48% ReqPass, 69.00% TaskPass, and 4.200 VQ. Compared with SCOPE, the strongest baseline overall, V-Protocol improves ReqPass by 12.70 percentage points and TaskPass by 27.00 percentage points. Direct generation, Mind-Brush, and GEMS complete substantially fewer tasks.

Table 2: **Commission completion on VCB.** Each tier reports REQPASS (RP), TASKPASS (TP), and visual quality (VQ). RP and TP are percentages. VQ averages 1–5 ratings, with missing or undecodable required images assigned zero. Overall reports benchmark-wide results. V-Protocol achieves the highest scores, while complete commissions remain challenging even when visual quality is high.

	S			M			C			Overall		
System	RP↑	TP↑	VQ↑	RP↑	TP↑	VQ↑	RP↑	TP↑	VQ↑	RP↑	TP↑	VQ↑
Direct	8.33	0.50	2.396	12.15	0.00	2.039	12.81	0.00	2.091	11.14	0.17	2.175
Mind-Brush	13.83	1.00	2.794	6.71	0.00	2.010	15.32	0.00	2.121	12.69	0.33	2.308
GEMS	41.20	16.00	3.639	10.49	0.00	2.236	34.69	8.00	2.337	30.92	8.00	2.737
SCOPE	84.36	55.00	3.658	77.74	39.50	4.097	82.10	31.50	4.192	81.78	42.00	3.982
V-Protocol	**96.95**	**82.00**	**4.250**	**91.78**	**67.00**	**4.157**	**94.15**	**58.00**	**4.195**	**94.48**	**69.00**	**4.200**

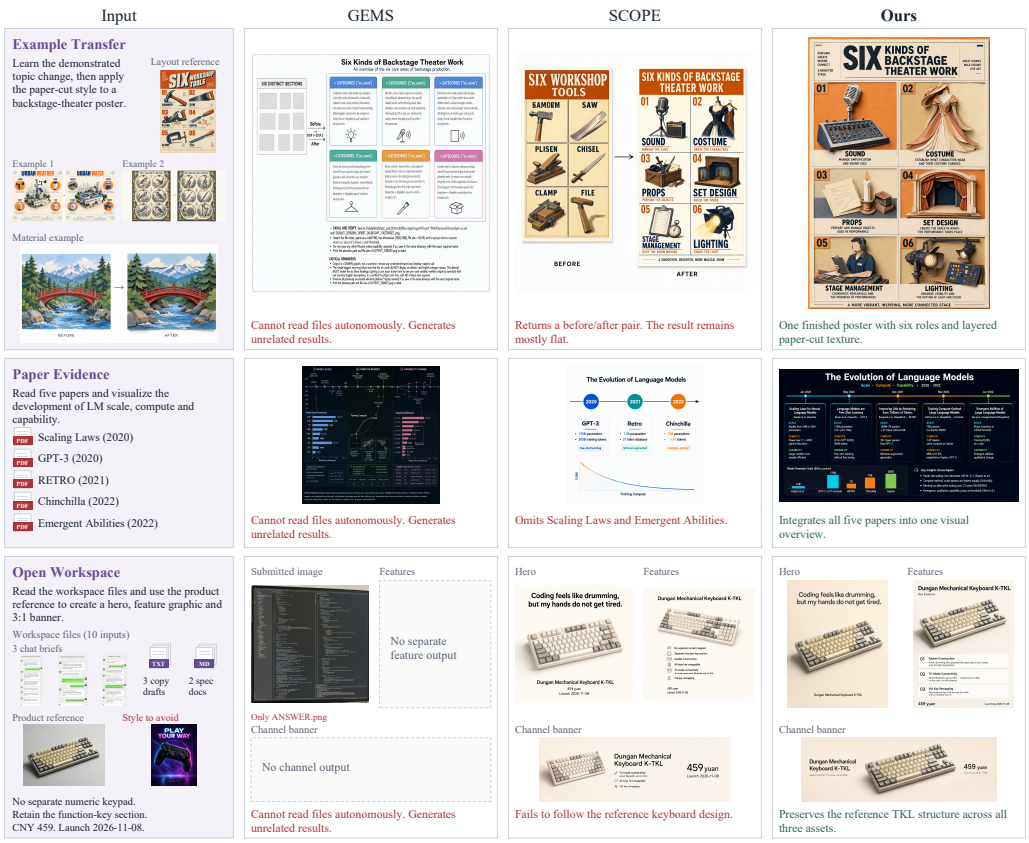

Figure 4: **Qualitative comparison.** V-Protocol achives better visual commissions completion.

Completion remains challenging as task scope expands. V-Protocol's TASKPASS decreases from 82.00% on S to 67.00% on M and 58.00% on C. SCOPE follows the same trend. V-Protocol's C-tier REQPASS of 94.15% also contrasts with its 58.00% TASKPASS, highlighting the remaining challenge of satisfying all requirements within a commission. Figure 4 complements the quantitative results with a qualitative comparison of visual outputs on selected VCB tasks.

Public benchmark for image agents. Table 3 shows that V-Protocol also achieves the highest overall scores among the compared systems on WISE-Verified and Mind-Bench. WISE-Verified

Table 3: **Results on WISE-Verified and Mind-Bench.** V-Protocol achieves the highest overall scores among the compared systems, with a larger advantage over SCOPE on reasoning-driven Mind-Bench tasks. WISE-Verified reports binary WiScore, and Mind-Bench reports CSA. Knowledge and Reasoning denote the corresponding category averages. The first three rows report published references; the remaining rows use shared backends. Bold marks the best results within the shared-backend group.

	WISE-Verified							Mind-Bench		
System	Culture	Time	Space	Biology	Physics	Chemistry	Overall	Knowledge	Reasoning	Overall
GPT-Image-1.5	0.8900	0.6917	0.8833	0.8000	0.7583	0.7750	0.8250	0.220	0.180	0.210
Nano Banana Pro	0.8975	0.8167	0.9333	0.8167	0.8667	0.8750	0.8760	0.400	0.440	0.410
Qwen-Image-Agent	0.9200	0.9167	0.9333	0.8333	0.8667	0.9000	0.9020	0.404	0.436	0.420
Direct	0.4800	0.4667	0.5583	0.3417	0.5417	0.3083	0.4580	0.032	0.052	0.042
Mind-Brush	0.6550	0.6417	0.6750	0.3750	0.5583	0.3500	0.5740	0.136	0.060	0.098
GEMS	0.8625	0.7833	0.9000	0.7667	0.7917	0.7167	0.8200	0.228	0.336	0.282
SCOPE	**0.9175**	**0.9333**	**0.9083**	0.8583	0.8500	0.9000	0.9010	**0.368**	0.416	0.392
V-Protocol	0.9100	0.9083	0.9000	**0.8750**	**0.9417**	**0.9167**	**0.9090**	**0.368**	**0.540**	**0.454**

Component analysis. We examine runtime enforcement and adaptive module selection on the same VCB subset. The runtime variant disables execution gates while retaining module selection and tool access. The selection variant replaces task-adaptive module choices with a predefined module set.

As shown in Table 4, the two variants achieve 65.00% and 66.67% TASKPASS, respectively, compared with 73.33% for the Full reference. Both also obtain lower REQPASS and VQ. These comparisons are exploratory because Full reuses historical outputs, including recovery runs, whereas the variants use a later frozen runtime. Detailed controls and additional results are provided in Appendix C.

Table 4: **Component analysis on VCB.**

Configuration	RP↑	TP↑	VQ↑
w/o Runtime Enforcement	84.46	65.00	3.769
w/o Adaptive Module Selection	80.25	66.67	3.553
Full V-Protocol	94.62	73.33	4.126

has a more focused task scope than VCB, and leading systems already perform well. V-Protocol scores 0.9090, narrowly ahead of SCOPE's 0.9010.

The difference is larger on Mind-Bench, where V-Protocol achieves 0.454 CSA compared with 0.392 for SCOPE. Their knowledge-driven averages are identical at 0.368, while V-Protocol improves the reasoning-driven average from 0.416 to 0.540. Its advantage on Mind-Bench is therefore concentrated in tasks requiring reasoning to determine the visual content. Together with the larger completion gap on VCB, these results show that comparable performance on knowledge-oriented generation does not necessarily imply comparable ability to complete visual commissions.

6 CONCLUSION

We introduced Visual Commission Bench (VCB) to evaluate image agents on complete visual commissions, from interpreting distributed source materials to delivering coordinated artifacts. V-Protocol supports these tasks by composing reusable task, cognitive, and expression modules through structured intermediate artifacts, while an execution harness tracks dependencies, revisions, and completion. With shared model backbones, V-Protocol achieves higher task success than the evaluated baselines. The commission-level results reveal substantial differences in completion despite comparable visual quality, demonstrating the importance of evaluating fulfillment of the entire request. Complete commissions remain challenging even with explicit execution support. Reliable use of intermediate results and preservation of requirements across revisions and deliverables remain priorities for further work.

7 AI USE STATEMENT

We used Gemini (Google) only for grammar correction and language polishing during the writing of this paper. No AI tools were involved in the research ideation, methodology, experiments, or analysis. The authors take full responsibility for all content in this manuscript.

REFERENCES

Harold Haodong Chen, Zhiyu Hou, Wen-Jie Shu, Weilin Ruan, Yingjie Xu, Litao Guo, and Ying-Cong Chen. Genrouter: Unified workflow routing for agentic image generation. *arXiv preprint arXiv:2608.16721*, 2026a.

Sixiang Chen, Zhaohu Xing, Tian Ye, Xinyu Geng, Yunlong Lin, Jianyu Lai, Xuanhua He, Fuxiang Zhai, Jialin Gao, and Lei Zhu. Genevolve: Self-evolving image generation agents via tool-orchestrated visual experience distillation. *arXiv preprint arXiv:2605.21605*, 2026b.

Jayanaka Dantanarayana, Savini Kashmira, Lingjia Tang, and Jason Mars. Sigil: Compiling agent skills into typed harnesses. *arXiv preprint arXiv:2607.27309*, 2026.

Gen Dong, Yanjie Gao, Liqun Li, Tianyin Xu, Yu Hua, and Fan Yang. Agent skills can be harmful: An empirical study of skill-induced failures in llm agents. *arXiv preprint arXiv:2608.11888*, 2026.

Kaituo Feng, Manyuan Zhang, Shuang Chen, Yunlong Lin, Kaixuan Fan, Yilei Jiang, Hongyu Li, Dian Zheng, Chenyang Wang, and Xiangyu Yue. Gen-searcher: Reinforcing agentic search for image generation. *arXiv preprint arXiv:2603.28767*, 2026.

Jun He, Junyan Ye, Zilong Huang, Dongzhi Jiang, Chenjue Zhang, Leqi Zhu, Renrui Zhang, Xiang Zhang, and Weijia Li. Mind-brush: Integrating agentic cognitive search and reasoning into image generation. *arXiv preprint arXiv:2602.01756*, 2026a.

Zefeng He, Siyuan Huang, Xiaoye Qu, Yafu Li, Tong Zhu, Yu Cheng, and Yang Yang. Gems: Agent-native multimodal generation with memory and skills. *arXiv preprint arXiv:2603.28088*, 2026b.

Dayeon Ki, Tianyi Zhou, Marine Carpuat, Gang Wu, Puneet Mathur, and Viswanathan Swaminathan. Graphicbench: A planning benchmark for graphic design with language agents. *arXiv preprint arXiv:2504.11571*, 2025.

Hao Li, Chunjiang Mu, Jianhao Chen, Siyue Ren, Zhiyao Cui, Yiqun Zhang, Lei Bai, and Shuyue Hu. Organizing, orchestrating, and benchmarking agent skills at ecosystem scale. *arXiv preprint arXiv:2603.02176*, 2026a.

Niantong Li, Guangzheng Hu, Weixu Qiao, Ying Ba, Qichen Hong, Shijun Shen, Jinlin Wang, Fan Zhou, Jianye Kang, Xin Shang, Ziyi He, Wei Wang, Dalin Li, Jiahao Li, Jie Zhang, Kaiyuan Gao, Kun Yan, Lihan Jiang, Ningyuan Tang, Shengming Yin, Tianhe Wu, Xiao Xu, Xiaoyue Chen, Yuxiang Chen, Yan Shu, Yanran Zhang, Yilei Chen, Yixian Xu, Zekai Zhang, Zhendong Wang, Zihao Liu, Zikai Zhou, Hongzhu Shi, Yi Wang, Bing Zhao, Hu Wei, Lin Qu, and Chenfei Wu. Qwen-image-bench: From generation to creation in text-to-image evaluation. *arXiv preprint arXiv:2605.28091*, 2026b.

Xiangyi Li, Yimin Liu, Wenbo Chen, Bingran You, Zonglin Di, Yifeng He, Shenghan Zheng, Kyoung Whan Choe, Jiankai Sun, Shuyi Wang, et al. Skillsbench: Benchmarking how well agent skills work across diverse tasks. *arXiv preprint arXiv:2602.12670*, 2026c.

Zongxia Li, Dawei Liu, Fuxiao Liu, Yuhang Zhou, Xiyang Wu, Jingxi Chen, Jing Xie, Xiaomin Wu, and Lichao Sun. Comfyclaw: Self-evolving skill harnesses for image generation workflows. *arXiv preprint arXiv:2607.01709*, 2026d.

Yujian Liu, Jiabao Ji, Li An, Tommi Jaakkola, Yang Zhang, and Shiyu Chang. How well do agentic skills work in the wild: Benchmarking llm skill usage in realistic settings. *arXiv preprint arXiv:2604.04323*, 2026.

Zijian Lu, Yiping Zuo, Yupeng Nie, Xin He, Weibei Fan, Lianyong Qi, and Shi Jin. Contractskill: Repairable contract-based skills for multimodal web agents. *arXiv preprint arXiv:2603.20340*, 2026.

Yuwei Niu, Munan Ning, Mengren Zheng, Weiyang Jin, Bin Lin, Peng Jin, Jiaqi Liao, Chaoran Feng, Fanqing Meng, Kun-Peng Ning, Bin Zhu, and Li Yuan. WISE: World knowledge-informed semantic evaluation for text-to-image generation. 2026.

Linyue Pan, Lexiao Zou, Shuo Guo, Jingchen Ni, and Hai-Tao Zheng. Natural-language agent harnesses. *arXiv preprint arXiv:2603.25723*, 2026.

Tianfei Ren, Zhipeng Yan, Yiming Zhao, Zhen Fang, Yu Zeng, Guohui Zhang, Hang Xu, Xiaoxiao Ma, Shiting Huang, Ke Xu, et al. Scope: Structured decomposition and conditional skill orchestration for complex image generation. *arXiv preprint arXiv:2605.08043*, 2026.

Boyin Tan, Xiaowei Huang, and Youcheng Sun. Skill coverage: A test adequacy metric for agent skills. *arXiv preprint arXiv:2606.20659*, 2026.

Yinghao Tang, Xueding Liu, Boyuan Zhang, Tingfeng Lan, Yupeng Xie, Jiale Lao, Yiyao Wang, Haoxuan Li, Tingting Gao, Bo Pan, et al. Igenbench: Benchmarking the reliability of text-to-infographic generation. *arXiv preprint arXiv:2601.04498*, 2026.

Yangbo Wei, Zhen Huang, Shaoqiang Lu, Junhong Qian, Qifan Wang, Chen Wu, and Lei He. Skillsmith: Co-evolving skills and tools for self-improving agent systems. *arXiv preprint arXiv:2606.01314*, 2026.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Tianle Xia, Lingxiang Hu, Yiding Sun, Ming Xu, Lan Xu, Siying Wang, Wei Xu, and Jie Jiang. Grasp: Graph-structured skill compositions for llm agents. *arXiv preprint arXiv:2604.17870*, 2026.

Junyan Ye, Jun He, Zilong Huang, Dongzhi Jiang, Xuan Yang, Rui Chen, and Weijia Li. Genclaw: Code-driven agentic image generation. *arXiv preprint arXiv:2605.30248*, 2026.

Zekai Zhang, Jiahao Li, Jie Zhang, Kaiyuan Gao, Kun Yan, Lihan Jiang, Ningyuan Tang, Shengming Yin, Tianhe Wu, Xiaoyue Chen, et al. Qwen-image-agent: Bridging the context gap in real-world image generation. *arXiv preprint arXiv:2606.26907*, 2026.

Jingyuan Zheng, Dongjing Wang, Xin Zhang, Butian Huang, Haiping Zhang, Dongjin Yu, and Shuguang Deng. Skillselect-serve: Budget-controllable and qos-aware skill service recommendation and composition for small llm agents. *arXiv preprint arXiv:2607.00011*, 2026.