ServImage: An Image Generation and Editing Benchmark from Real-world Commercial Imaging Services

Fengxian Ji[1]*, **Jingpu Yang**[2]*, **Zirui Song**[1]*, **Lang Gao**[1], **Junhong Liang**[1], **Zhenhao Chen**[1], **Jinghui Zhang**[1], **Xiuying Chen**[1]†
[1]MBZUAI, United Arab Emirates
[2]Institute of Botany, the Chinese Academy of Sciences, China
{fengxian.ji, zirui.song, lang.gao junhong.liang zhenhao.chen }@mbzuai.ac.ae
{jinghui.zhang, xiuying.chen}@mbzuai.ac.ae
{jingpuyang290}@gmail.com

Abstract

Recent image generation and editing models demonstrate robust adherence to instructions and high visual quality on academic benchmarks. However, their performance on paid, real-world design projects remains uncertain. We introduce **ServImage**, a benchmark that explicitly correlates model outputs with economic value in commercial design projects. ServImage consists of (i) ***ServImageBench***: a dataset of 1.07k paid commercial design tasks and 2.05k designer deliverables totaling over $295k, covering portrait, product, and digital content, along with 33k candidate images and 33k human annotations. (ii) ***ServImageScore***: an integrated scoring system that combines three quality dimensions: baseline requirements fulfilment, visual execution quality, and commercial necessity satisfaction. These three dimensions are designed to characterize the factors that drive human payment decisions and indicate whether an image is commercially acceptable. (iii) ***ServImageModel***: under this scoring system, we propose a payment prediction model trained on the human-annotated candidate images, achieving 82.00% accuracy in predicting human payment decisions and producing calibrated payment probabilities. ServImage provides a comprehensive foundation for assessing the commercial viability of image generation models and offers a scalable resource for future research on economically grounded vision systems Github.

1 Introduction

Recent years have witnessed remarkable progress in image generation and editing. In image generation, diffusion models such as Stable Diffusion (Rombach et al., 2022) and GLIDE (Nichol et al., 2021) have advanced high-quality text-to-image generation. In image editing, models like InstructPix2Pix (Brooks et al., 2023) and MagicBrush (Zhang et al., 2023, 2025a) have improved editing precision and controllability. These advances have broadened the use of image generation from conventional tasks such as object removal and style transfer to more complex design tasks, including poster design, logo creation, and IP illustration (Chen et al., 2024; Wang et al., 2025b), which originate from real-world commercial scenarios with well-defined requirements and economic returns. On the evaluation side, recent benchmarks emphasized instruction following and semantic controllability (Ma et al., 2024), optical realism and reflection consistency (Zeng et al., 2024), and physical commonsense understanding, ensuring that generated images obey plausible geometry, gravity, and material properties (Farshad et al., 2023; Zhao et al., 2025; Ryu et al., 2025; Gu et al., 2025). However, to the best of our knowledge, no existing benchmark evaluates how well image models satisfy real-world commercial design requirements, particularly in terms of practical usability and economic acceptability.

In real design services, commercial acceptability depends on a coupled set of factors beyond semantic and perceptual correctness. A deliverable must first satisfy baseline business requirements, such as size, resolution, format, copyright safety, and policy compliance, before it can even be considered for delivery. It must then achieve sufficient visual execution quality to be publishable, and, crucially, align with the client's specific commercial intent, including brand style, marketing message, and platform constraints. Whether a client approves and pays for an image ultimately reflects these business-rule, visual, and commercial considerations jointly, which are not explicitly captured by existing benchmarks. The fundamental question in a business context is not merely "Is it good?" but rather "Is it worth paying for?" (Patwardhan et al., 2025; Mazeika et al., 2025).

*Equal contribution.
†Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 43504–43529
July 2-7, 2026

Figure 1: Overview of the ServImage benchmark and evaluation framework. (a) We collect 1,070 paid design tasks from online crowdsourcing platforms and group them into Portrait, Product, and Digital categories. (b) Through the dataset pipeline with rule extraction/point mining, we obtain BRF, VEQ, and CNS scores to form ServImageScore. (c) Sixteen image generation and editing models produce about 33k candidate images. (d) Optionally, built on these scores, ServImageModel predicts human payment decisions. (e) We present representative tasks and model outputs as ServImage case studies. (f) Under the standard settlement scenario, we evaluate each model independently on the full task set and compare their commercial capability using the economic metric of total revenue.

To close this gap, we introduce *ServImage*, a benchmark explicitly grounded in real monetary outcomes for commercial design services. Concretely, *ServImage* contains three components. ***ServImageBench*** collects 1.07k paid commercial design tasks from crowdsourcing platforms, with over $295k of contract value, together with 33k candidate images and 33k human payment decisions from 16 mainstream models, spanning portrait services, e-commerce products, and digital content, and covering both text-to-image generation and image-editing workflows. ***ServImageScore*** provides an integrated evaluation scheme that decomposes image quality into Baseline Requirements Fulfilment, Visual Execution Quality, and Commercial Necessity Satisfaction, with a unified scoring protocol that jointly captures business-rule adherence, perceptual quality, and task-specific commercial fit. Finally, ***ServImageModel*** is a settlement model that, guided by ServImageScore and task prices, produces calibrated payment probabilities, serving as a reference method for estimating the expected revenue of different models under realistic payment rules. ServImage demonstrates from the perspective of commercial payment that models which appear strong under technical metrics capture only a fraction of the attainable commercial value.

In summary, our contributions include: advancing generative model evaluation from traditional technical metrics to a market-grounded perspective based on real paid tasks, human payment behavior, and price structures; uncovering the mechanisms by which commercial value is formed, showing that user preferences, task prices jointly determine economic outcomes beyond image quality alone; and providing an economic inference framework that estimates attainable revenue under realistic payment rules, revealing a substantial gap between technical performance and actual commercial value capture.

2 Related Work

Image Generation and Editing Benchmarks. Evaluation of generative models has progressed from traditional single-number perceptual metrics, CLIP Score, LPIPS, and PSNR/SSIM (Radford et al., 2021; Zhang et al., 2018; Wang et al., 2004) to multi-dimensional assessment frameworks. Generation-oriented benchmarks such as GenAI-Bench and OmniGenBench foreground compositional reasoning and text rendering (Li et al., 2024; Peng et al., 2024; Zhou et al., 2025),

while editing-oriented ones like EDITVAL, VIEScore, and IE-Bench emphasise controllability and region consistency (Basu et al., 2023; Sun et al., 2025; Wang et al., 2025a). Unified efforts, including I2EBench, ICE-Bench, and RISEBench, attempt to bridge this divide by jointly evaluating instruction following, visual quality, and reasoning (Ma et al., 2024; Pan et al., 2025; Xu et al., 2025b). However, a fundamental limitation persists: existing tasks lack market-validated value distributions, and current evaluations rely on conventional technical metrics that provide no insight into commercial viability.

VLM-based evaluation for image generation and edition. To automate assessment, the VLM-as-a-Judge paradigm has become prevalent, using task-specific prompts to achieve high correlations with human preference (Wu et al., 2025; Ye et al., 2025; Pu et al., 2025; Wang et al., 2025b; Zhang et al., 2025b; Wang et al.; Yang et al., 2026). Systems like LMM4Edit (Xu et al., 2025a; Ji et al., 2025) use few-shot learning for robust editing evaluation. While these methods align well with human perception, they don't capture the economic value of outputs. This gap shows a decoupling where high technical metrics don't necessarily lead to willingness to pay. Although similar mappings from model performance to economic outcomes have been explored in other domains, such as SWE-Lancer for code generation (Miserendino et al., 2025; Li et al., 2026; Yang et al., 2025; Liang and Zhang, 2025), the image domain still lacks a verifiable, monetarily grounded framework. Detailed comparison is in Appendix Table 5.

3 ServImage

3.1 Task Formulation

We begin by formalizing the structure of ServImage tasks and the associated payment contracts. Let $\mathcal{T}$ denote the set of paid design tasks in ServImage. For each task $t \in \mathcal{T}$, we denote the task context by $x_t = (\text{brief}_t, \text{refs}_t, \text{src}_t)$, where brief_t is a textual description of the design goal, refs_t are optional reference materials (e.g., sketches, logos, exemplar images), and src_t is an optional source image when the client requests editing rather than pure generation. Each task further specifies a contract price $\text{Price}(t)$ and a required number of deliverables $Q(t)$. We define the implied per-deliverable Price as $p_{\text{img}}(t) = \text{Price}(t)/Q(t)$.

Category	#Tasks	#Deliverables	Total Value	Avg Price
Portrait	271	288	\$15.3k	\$56.41
Product	450	839	\$147.0k	\$326.65
Digital	349	919	\$132.8k	\$380.40
All	1,070	2,046	\$295.0k	\$275.74

Table 1: Overall statistics of the ServImage benchmark.

Given a task context x_t, a model produces a set of candidate images $\hat{Y}_t = \{\hat{\text{img}}_{t,i}\}_{i=1}^{Q(t)}$. Each generated image is assigned a binary validity label ($S_{t,i} \in 0, 1$) from human payment decisions, indicating whether it is accepted as a deliverable. The total value earned by the model on task t is then:

$$V_t(\hat{Y}_t) = \sum_{i=1}^{Q(t)} S_{t,i}\, p_{\text{img}}(t), \tag{1}$$

which measures how much the model would earn on task t under the same payment contract as human designers.

3.2 ServImageBench

With this task formulation in place, we now describe ServImageBench, the real-world dataset from which these tasks and payment contracts are instantiated. ServImageBench is constructed from paid design orders collected between 2018 and 2025 from two major Chinese online crowdsourcing platforms, Epwk.com and Zbj.com. Each raw posting specifies a concrete design goal, a quoted budget, and delivery requirements. In these orders, clients typically provide a textual brief and optional reference materials, expecting commercially usable visual assets tailored to a specific use case. After cleaning, ServImage consists of 1,070 paid commercial design tasks and 2,046 deliverables, with summary statistics reported in Table 1. Across the benchmark, 833 single-image tasks account for 77.9% of the total, while the remaining 236 are multi-image tasks representing 22.1%. We consider a task as editing if its brief includes at least one input image; otherwise, it is a pure text-to-image creation. Under this definition, 74.8% of tasks are text-to-image creation and 25.2% involve editing an existing image. Representative task examples from the three categories are shown in Appendix Sec C.

All tasks are grouped into three client-driven commercial categories that correspond to major freelance design verticals: Portrait, Product, and Digital. The Portrait category covers personal imaging services such as ID photo retouching, profile pictures, and customized character portraits

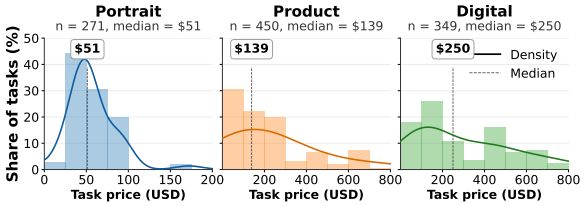


Figure 2: Task price distributions for Portrait, Product, and Digital categories. Dashed lines indicate median prices, showing a long-tailed pattern across categories.

for social media or branding, and is dominated by single-image jobs, reflecting the one-off nature of personal imaging orders. Product tasks focus on e-commerce and commercial product visuals, including logo design, brand identity systems, product posters, and packaging layouts for online stores and marketing campaigns. Digital tasks involve broader digital content and web-oriented visual assets, such as UI mockups, web banners, illustrations, IP characters, and various online media creatives for websites, apps, or social platforms. Table 1 and Figure 2 summarize the number of tasks and deliverables, as well as the total contract value across these categories.

3.3 ServImageScore

Existing benchmarks for image generation models rarely measure their economic value: their metrics mainly focus on instruction following and visual fidelity, which are not directly applicable to commercial design services. Starting from how real clients judge deliverables, we define three practical aspects and bundle them into a unified label tuple, **ServImageScore**, which represents commercial acceptability through BRF, VEQ, and CNS together with the deliverable-level binary payment label, rather than through a fixed weighted aggregation. For each task t with associated images i, ServImageScore consists of a baseline requirement fulfillment score at task-level $\mathrm{BRF}(t) \in [0,5]$ shared by all images in t, and two image-level scores $\mathrm{VEQ}(t,i), \mathrm{CNS}(t,i) \in [0,5]$ defined for each image (t,i), together with a binary payment decision $\mathrm{Accept}(t,i) \in \{0,1\}$ indicating whether the deliverable is labeled as approved and paid under the original task contract by human annotators.

Baseline Requirements Fulfilment, BRF. BRF measures whether a candidate satisfies the explicit requirements in the client brief, such as logo placement and background colour. We first run a rule-extraction pipeline that decomposes each brief into binary evaluation points, with the VLM predicting completion (0/1) for each by comparing the image against the brief and references. For multi-image tasks, an evaluation point is satisfied if at least one image meets it. We then compute the completion rate $c_t = \frac{1}{K_t}\sum_{k=1}^{K_t} r_{t,k} \in [0,1]$, which represents the fraction of requirements satisfied for task t. To keep BRF on the same $[0,5]$ scale as the other ServImageScore components, we linearly rescale this rate and define the task-level score as $\mathrm{BRF}(t) = 5c_t \in [0,5]$, and this value is shared by all images i in task t. For notational uniformity at the image level, we define $\mathrm{BRF}(t,i) := \mathrm{BRF}(t)$ for all images i in task t.

Visual Execution Quality, VEQ. VEQ assesses the overall visual quality of an image, independent of whether it follows the brief. For each image i in task t, the VLM is instructed to consider three aspects: (i) technical quality (clarity, artefacts, realism), (ii) aesthetic quality (composition, colour, lighting, harmony), and (iii) text quality when textual elements are present (readability and typography). Based on this rubric, the VLM directly outputs a single visual-quality score $\mathrm{VEQ}(t,i) \in [0,5]$ for image (t,i). When an image contains no textual elements, the text-quality aspect is marked as N/A and excluded from aggregation, and $\mathrm{VEQ}(t,i)$ is computed from the remaining aspects but still lies on the same $[0,5]$ scale.

Commercial Necessity Satisfaction, CNS. CNS captures consistency in settings where relationships between images matter, namely, image editing and multi-image tasks. For editing tasks, $\mathrm{CNS}(t,i)$ assesses the preservation of non-edited regions, key subject attributes, natural boundaries, and coherent lighting and perspective. For multi-image tasks, we compute a set-level score $\mathrm{CNS}(t)$ for cross-image consistency in style, colour palette, and brand elements, and assign $\mathrm{CNS}(t,i) := \mathrm{CNS}(t)$ to all images in the task; when a task does not involve editing or multiple images, $\mathrm{CNS}(t,i)$ is marked as N/A and excluded from score aggregation.

3.4 Data Annotation

To operationalise the ServImageScore framework and obtain ground-truth labelsfor later modelling Sec. 4, we annotate model-generated candidates across all tasks in ServImageBench. Concretely, we run the 16 image generation and editing models introduced in Sec. 5, denoted by $\mathcal{M}$, on all

paid tasks. For each task $t \in \mathcal{T}$ and model $m \in \mathcal{M}$, we generate $Q(t)$ candidate deliverables $\{\hat{\text{img}}_{t,i,m}\}_{i=1}^{Q(t)}$ using the task brief, reference materials, and source image if provided. Collectively, these outputs form an extended candidate set $\mathcal{D}_{33\text{K}}$ of approximately 33k images.

Each candidate $(t, i, m) \in \mathcal{D}_{33\text{K}}$ inherits the original task context and is annotated with the ServImageScore tuple defined in Sec. 3.3. That is, we obtain a task-level BRF score $\text{BRF}(t)$ and image-level scores $\text{VEQ}(t, i)$ and $\text{CNS}(t, i)$, all in $[0, 5]$, together with a binary payment label $y_{t,i} \in \{0, 1\}$, where $y_{t,i} = 1$ indicates that the candidate would be approved and paid under the original task contract, and $y_{t,i} = 0$ otherwise. In all experiments, we treat $y_{t,i}$ as the ground-truth settlement outcome; ServImageModel is used only as an auxiliary predictor for expected-revenue estimation. The triple $(\text{BRF}(t, i), \text{VEQ}(t, i), \text{CNS}(t, i))$ serves as the ground-truth concept scores for Sec. 4, while $y_{t,i}$ is the ground-truth payment decision. As shown in Fig. 3, a composite score derived from the annotated concepts $\text{s}_{t,i}$ is monotonically correlated with empirical acceptance rates, supporting these dimensions as a meaningful concept space for payment prediction.

The BRF, VEQ, and CNS scores are obtained automatically through a unified VLM-as-a-judge pipeline described in Sec. 5, whereas the payment labels $y_{t,i}$ are assigned by trained human annotators following platform-style guidelines. The full prompts are provided in Appendix E. We adopt a double-annotation-with-adjudication protocol, following prior work on expert-labelled datasets (Jin et al., 2019). Each deliverable is independently labelled by two annotators. Across 2,000 randomly sampled doubly annotated instances, we observe an inter-annotator agreement of 67.92%. Any remaining disagreements are resolved by a third annotator, who adjudicates the final label.

4 ServImageModel

With the annotated candidate dataset in place, we now turn to modelling human payment decisions. Our objective is to train ServImageModel, a predictor of whether each candidate image $\hat{\text{img}}_{t,i}$ in task t would be accepted and paid according to human payment labels $y_{t,i}$. Concretely, ServImageModel is a two-stage neural network consisting of (i) a concept predictor and (ii) a payment head f_θ. Given the task context x_t and a candidate image $\hat{\text{img}}_{t,i}$, the model first predicts the three ServImageScore dimensions as intermediate concepts, and then uses these predicted concepts to estimate the final acceptance probability.

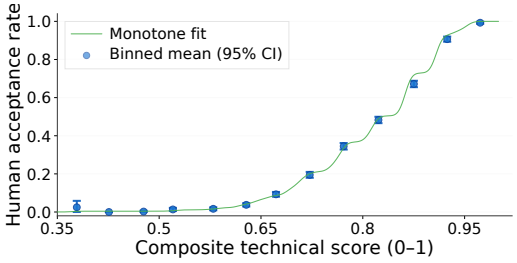


Figure 3: Composite scores from BRF, VEQ, and CNS correlate with acceptance rates on ServImage-33K, showing that $\mathbf{s}_{t,i}$ aids payment prediction. Data splits are at the task level to prevent leakage across deliverables from the same order.

(1) Concept-bottleneck prediction from the three quality dimensions. In our case, the three ServImageScore dimensions act as these explicit intermediate concepts. For each candidate (t, i), we denote the annotated concept vector as $\text{s}_{t,i} = (\text{BRF}(t, i), \text{VEQ}(t, i), \text{CNS}(t, i))$. During training, the model predicts a corresponding concept vector $\hat{\text{s}}_{t,i}$, and matches it to the annotated scores using regression losses $\mathcal{L}_{\text{BRF}}^{(n)}$, $\mathcal{L}_{\text{VEQ}}^{(n)}$, and $\mathcal{L}_{\text{CNS}}^{(n)}$. The concept loss is as:

$$\mathcal{L}_{\text{CBM}} = \frac{1}{N} \sum_{n=1}^{N} \left(\mathcal{L}_{\text{BRF}}^{(n)} + \mathcal{L}_{\text{VEQ}}^{(n)} + \mathcal{L}_{\text{CNS}}^{(n)} \right). \quad (2)$$

(2) Payment prediction with a task-aware head. On top of the learned concepts, we train a payment predictor f_θ that takes the task context x_t, candidate image $\hat{\text{img}}_{t,i}$, and concept scores $\hat{\text{s}}_{t,i}$ as input and outputs an acceptance probability:

$$\hat{p}_{t,i} = f_\theta(\text{x}_t, \hat{\text{img}}_{t,i}, \hat{\text{s}}_{t,i}) \in [0, 1]. \quad (3)$$

The payment head is supervised using the binary payment labels $y_{t,i} \in \{0, 1\}$ obtained during annotation:

$$\mathcal{L}_{\text{pay}} = \mathbb{E}_{t,i}[-y_{t,i} \log \hat{p}_{t,i} - (1 - y_{t,i}) \log(1 - \hat{p}_{t,i})], \quad (4)$$

which encourages the predicted acceptance probability to align with empirical human payment decisions.

Concretely, we treat $\hat{p}_{t,i}$ as the acceptance probability of candidate image i for task t, aggregate these probabilities within each task to obtain a task-level acceptance rate, and then combine this rate with the task price to compute each model's total payment under our payment model.

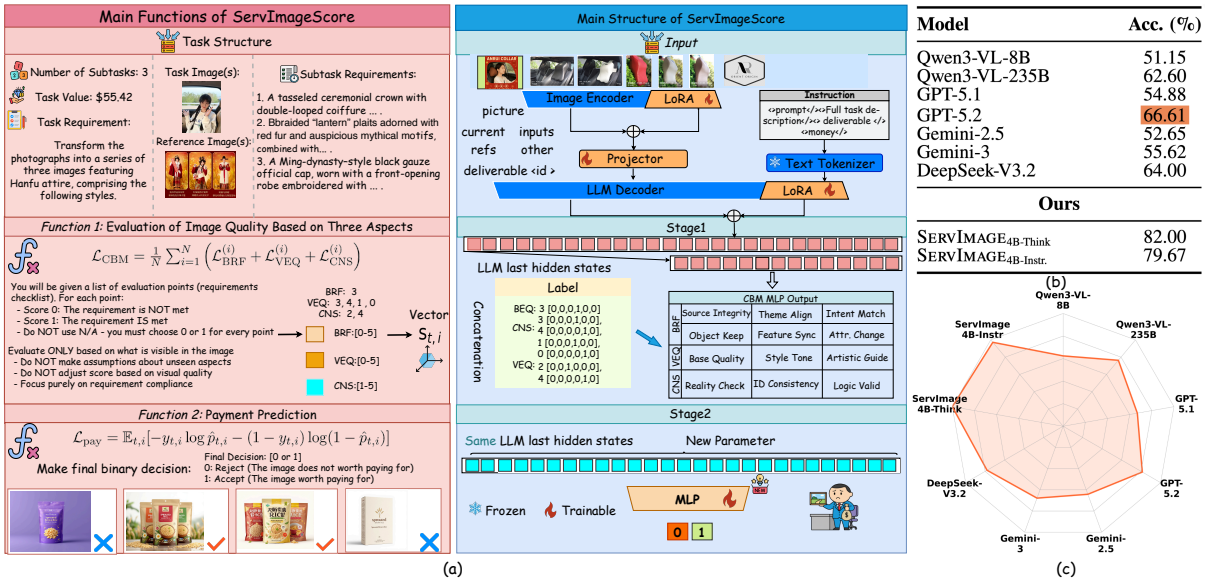


Figure 4: Overview of ServImageModel: (a) Two-stage ServImageModel architecture; (b) Accuracy comparison of base models and our variants. (c) Radar chart of Accuracy Comparison Result.

5 Experiments and Analysis

5.1 Setups

We evaluate 16 image models, including 12 proprietary and 4 open-source models, for text-to-image generation and image editing. On the commercial side, we include OpenAI's GPT-5-Image family, including gpt-5-image and gpt-5-image-mini (OpenAI, 2025b,a); Google's Gemini Nano Banana, Banana Pro, and Imagen-4.0 (DeepMind, 2025; Cloud, 2025); multiple Doubao systems, including Seedream 4.0 and Seedream 3.0-t2i for text-to-image generation (Seedream et al., 2025; Gao et al., 2025), and SeedEdit 3.0-i2i for image-to-image editing (Wang et al., 2025c); Kuaishou's Kling-v2 (Technology, 2025); and models from professional design platforms such as Ideogram-v3 (Ideogram, 2025) and Midjourney's Imagine and Edit pipelines (Midjourney, 2024). We also include the MJ-Relax-Edits and MJ-Relax-Imagine modes from Midjourney. On the open-source side, we include FLUX-1.1-Pro (Labs, 2024) and FLUX.1-Kontext-Pro (Labs, 2025), Qwen-Image (et al., 2025), and Stable-Diffusion-3.5-Large (AI, 2024). We use OpenAI's GPT-5-mini as an automatic judge to compute the BRF, VEQ, and CNS scores defined in Sec. 3.3.

5.2 Main Experiments

Under the standard settlement scenario, we evaluate 16 image models on ServImage and report their economic outcomes in Table 2. Overall, the top proprietary models capture most of the economic value, for example, Gemini-Banana-Pro earns $243.98k (82.7% share) and GPT-Image-1 earns $148.32k (50.3%), while the strongest open-source baseline, Qwen-Image, obtains a markedly smaller payout of $75.50k (25.6%). The best-performing model, Gemini-Banana-Pro, achieves the highest total revenue of $243.98k and captures the largest contract share of 82.7% out of the overall $295k budget. In contrast, the strongest open-source baseline, Qwen-Image, earns $75.50k with a 25.6% share, indicating a substantial gap in both maximum attainable revenue and market share under the same settlement rule.

Across the three fine-grained categories in ServImage, we observe clear leaders in economic outcomes. Kling-v2 leads Portrait with $9.79k revenue (64.0% within-category share), narrowly followed by MJ-Relax-Imagine (57.7%). Gemini-Banana-Pro ranks first in Product with $122.60k (83.4%), with GPT-Image-1-Mini second (59.8%), and also dominates Digital with $117.48k (88.5%), followed by Ideogram-v3 (44.1%). Despite these leaders, no single model achieves consistently high share across all categories, suggesting substantial room for improvement. Table 2 uses human accept/reject labels as ground truth; consequently, Table 9 reports an automatic proxy estimated from ServImageModel-predicted payment probabilities. Additionally, the failure examples and corresponding analysis are provided in Appendix I.

Model	Total				Portrait		Product		Digital	
	Rev ($k)	Share (%)	Task Acc. (%)	Deliv. Acc. (%)	Rev ($k)	Share (%)	Rev ($k)	Share (%)	Rev ($k)	Share (%)
Closed-Source Models										
Gemini-Banana	109.27	37.0	29.91	34.56	2.08	13.6	50.72	34.5	56.48	42.5
Gemini-Banana-Pro	243.98	82.7	70.50	79.88	3.90	25.5	122.60	83.4	117.48	88.5
GPT-Image-1	148.32	50.3	48.22	46.38	7.38	48.3	82.95	56.4	57.99	43.7
GPT-Image-1-Mini	113.04	38.3	32.36	37.62	0.17	1.1	87.95	59.8	24.93	18.8
Ideogram-v3	80.53	27.3	25.89	31.67	1.01	6.6	20.94	14.2	58.57	44.1
Imagen-4.0	111.58	37.8	31.03	28.10	4.03	26.3	70.46	47.9	37.09	27.9
Kling-v2	96.15	32.6	40.22	28.85	9.79	64.0	39.33	26.8	47.02	35.4
MJ-Relax-Edits	78.81	26.7	22.43	25.90	1.10	7.2	30.93	21.0	46.78	35.2
MJ-Relax-Imagine	88.08	29.9	37.66	28.01	8.82	57.7	48.20	32.8	31.06	23.4
SeedEdit-3.0-i2i	1.81	0.6	6.98	6.77	0.00	0.0	0.67	0.5	1.14	0.9
Seedream-3.0-t2i	105.21	35.7	31.40	41.54	0.95	6.2	71.46	48.6	32.80	24.7
Seedream-4.0	65.95	22.4	31.12	40.18	2.44	15.9	46.45	31.6	17.06	12.8
Open-Source Models										
FLUX-1.1-Pro	30.53	10.3	10.00	8.85	0.33	2.2	3.30	2.2	26.91	20.3
FLUX-Kontext-Pro	53.14	18.0	16.47	14.18	0.21	1.4	20.42	13.9	32.51	24.5
Qwen-Image	75.50	25.6	24.95	30.40	1.73	11.3	31.58	21.5	42.19	31.8
SD-3.5-Large	16.99	5.8	8.60	6.40	2.33	15.2	5.33	3.6	9.33	7.0

Table 2: Model performance on ServImage under the standard settlement scenario. Revenue, Rev ($k), is the total contract value earned by a model, and Share is its fraction of the overall contract value ($295k), while category shares are computed within each category. Task Acceptance and Deliverable Acceptance are the proportions of tasks and deliverables approved according to human payment decisions ($y_{t,i}$). The last six columns report revenue and share by Portrait, Product, and Digital categories. All monetary values are in thousand USD. **Color:** 1st 2nd 3rd.

Model	Rev ($k)	Rank
Gemini 2.5 Flash	78.9	2
Gemini-Banana-Pro	113.7	1
GPT-5 Image	10.1	5
GPT-5 Image Mini	5.8	6
imagen-4.0	4.1	7
Seedream 4.0	0.0	15
Seedream 3.0-t2i	0.3	13
SeedEdit 3.0-i2i	0.0	16
Kling-v2	3.2	9
ideogram-v3	0.9	10
MJ-Relax-Imagine	3.5	8
MJ-Relax-Edits	0.4	11
Open-Source Models		
stable-diffusion-3.5-large	0.1	14
Flux Pro	35.7	4
flux-kontext-pro	37.9	3
Qwen-Image	0.3	12

Table 3: Crowdsourcing competition earnings with winner-takes-all acceptance. We report total value earned Rev which means Revenue, and Rank. **Color:** 1st 2nd 3rd.

5.3 Crowdsourcing Competition

The settlement scenario above assumes that a single provider is deployed at a time: each model is evaluated on the full test split and earns revenue from all tasks it solves. In practice, commercial crowdsourcing platforms list a task to many providers, and only the best submission is paid. To capture this more competitive regime, we construct a crowdsourcing competition setting in which all models are run on the same tasks, and for each task, we select the winner by the number of deliverables accepted by human payment labels, ties are broken by the summed ServImageScore (BRF/VEQ/CNS). The full task revenue is then assigned to this "winner" model and all other models receive zero payment for that task, resulting in a winner-takes-all allocation of the total contract value, as shown in Table 3. In this section, a task is considered successful if and only if all its required deliverables are accepted (paid) according to the human payment-decision labels.

In the crowdsourcing competition setting, earnings are highly concentrated: a small number of top models capture a much larger portion of the total revenue, while many others earn little or even zero. However, the outcome is not completely one-sided. For example, Gemini-Banana-Pro ranks first with $113.7k, but Gemini 2.5 Flash remains close behind at $78.9k, and an open-source model (flux-kontext-pro) also achieves a strong result with $37.9k.

5.4 Cost Reduction Analysis

The revenue analysis in Sec. 5.2 quantifies each model's earnings under our settlement rule. In practice, platform operators and clients are also interested in a complementary question. If a model is deployed as the first stage of the pipeline and human designers only redo failed tasks, how much outsourcing cost can be saved, and how efficient is the model-first pipeline in terms of return per dollar spent?

We capture this economic perspective using model-level aggregate quan-

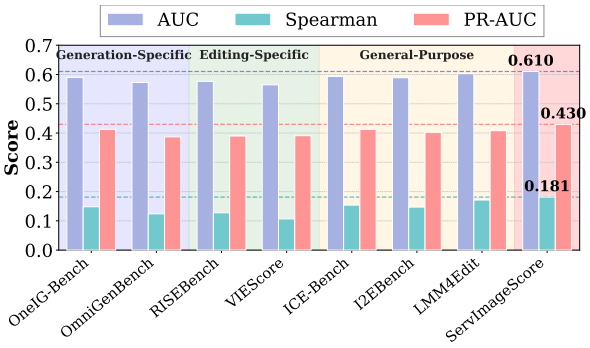


Figure 5: Metric comparison on the test set. Bars show AUC, Spearman, and PR-AUC for seven metrics and ServImageScore, higher is better.

tities. For a given model m, we define $B = \sum_{t\in\mathcal{T}} \textbf{Price}(t)$, $S_c(m) = 1 - \frac{\textbf{Cost}_{\text{API}}(m)}{B} - \frac{1}{B}\sum_{t\in\mathcal{T}} \textbf{Price}(t)(1 - \textbf{Success}_c(m,t))$, and the *Contribution Ratio* $R_c(m) = \frac{\sum_{t\in\mathcal{T}} \textbf{Price}(t)\,\textbf{Success}_c(m,t)}{\textbf{Cost}_{\text{API}}(m) + \sum_{t\in\mathcal{T}} \textbf{Price}(t)(1 - \textbf{Success}_c(m,t))}$. Here $S_c(m)$ is the overall cost savings relative to the human-only baseline, while $R_c(m)$ measures the return per dollar spent on the model API and freelancer rework under scenario c. In the main text we report them as Cost Savings (%) and Contribution Ratio. Per-call API price assumptions are listed in Appendix Table 6.

Empirically, the best-performing model, MJ-Relax-Edits, achieves cost savings of $S_c(m) \approx 81.7\%$ and a contribution ratio of $R_c(m) \approx$, indicating that each dollar spent on API usage and human rework yields about $4.4581 of end-to-end contract value handled by the model-first pipeline under realistic settlement rules. Appendix Table 7 reports the full cost reduction results.

5.5 Evaluation Method Comparison

To evaluate ServImageScore in a payment-decision setting, we compare it with seven metrics in three categories. Generation-specific metrics (OmniGenBench (Wang et al., 2025b), OneIG-Bench (Chang et al., 2025)) focus on text-to-image and do not fit source-constrained editing. Editing-specific metrics (VIEScore (Ye et al., 2025), RISEBench (Zhao et al., 2025)) target editing but not source-free generation. General-purpose metrics include ICE-Bench (Pan et al., 2025), I2EBench (Ma et al., 2024), and LMM4Edit.

We examine whether automatic metrics can approximate human payment decisions. We apply each metric to the same subset and evaluate its alignment with human payment labels using three criteria: (i) AUC, the area under the ROC curve for binary payment outcomes; (ii) Spearman, the rank correlation between metric scores and payment labels; and (iii) PR-AUC, the area under the precision–recall curve for payment prediction. Across all three metric groups, existing baselines show lower Spearman and PR-AUC than ServImageScore, suggesting weaker alignment with commercial acceptance. As shown in Fig. 5, ServImageScore performs best on all three criteria.

Additionally, the ServImageScore of different models, together with the judge robustness and reproducibility results relative to human evaluation, are provided in Appendix H, where the analysis shows that ServImageScore is robust.

5.6 ServImageModel Analysis

Comparison with General-purpose VLMs. Fig. 4(b) compares the payment-decision accuracy (%) of general-purpose VLM baselines and ServImageModel on the test split. Among all general-purpose baselines, our task-specific models substantially outperform all baselines: ServImageModel$_{\text{4B-Think}}$ reaches 82.00%.

Base Model	CBM	LoRA (v)	LoRA (l)	Acc.
Qwen-4B-Think				72.12%
Qwen-4B-Think	✓			78.30%
Qwen-4B-Think	✓	✓		79.21%
Qwen-4B-Think	✓	✓	✓	**82.00%**
Qwen-4B-Instr.	✓	✓	✓	79.67%

Table 4: Ablation study on the CBM interface, LoRA strategies, and Qwen3-VL backbones. We report payment-decision accuracy on the test split.

Ablation Study. To further understand why ServImageModel. achieves large gains in Fig. 4(b), we ablate three components of the payment model: the CBM interface, LoRA adapters on the vision and language towers, and the choice of Qwen3-VL backbone, as shown in Table 4. Using Qwen-4B-Think as the base model, the plain baseline achieves 72.12% accuracy. Adding LoRA on both vision and language yields the best performance of 82.00%. Together, these two findings demonstrate the effectiveness of the ServImageModel framework for predicting payment decisions.

6 Conclusion

We present ServImage, a benchmark for evaluating image generation and editing models on real-

world commercial design tasks and payment decisions. By connecting model outputs with task prices and human acceptance outcomes, and introducing ServImageScore to decompose commercial acceptability into baseline requirement fulfillment, visual execution quality, and commercial necessity satisfaction, ServImage moves beyond conventional metrics and better reflects why people pay. We hope ServImage will encourage market-aware evaluation and foster vision models that are more economically aligned.

Limitations

ServImageBench is constructed from paid design tasks collected from two Chinese crowdsourcing platforms. As a result, task distributions, aesthetic preferences, and pricing norms may vary across regions, languages, and procurement settings. As a first look at linking image generation and editing performance to real commercial payment outcomes, we view this benchmark as an exploratory starting point, and future efforts can broaden coverage to more regions, languages, and enterprise-level design scenarios to improve generalizability.

In addition, Commercial acceptability is currently operationalized as a binary payment decision, which provides a clear and reproducible economic signal consistent with the settlement rules of real crowdsourcing platforms. However, real-world design workflows are typically more iterative and fine-grained, involving partial acceptance, repeated revisions, and negotiated outcomes. Under such conditions, commercial value depends not only on initial output quality but also on a model's stability, editability, and robustness across multiple rounds of interaction. Consequently, the current benchmark does not fully capture the iterative nature of commercial design or its broader economic implications. Future work may extend this formulation by incorporating revision-based workflows, multi-turn evaluation, and more continuous representations of commercial outcomes.

Finally, Prior public evaluations largely use human preference signals or subjective ratings, with emerging efforts incorporating online utility metrics such as engagement or CTR. In contrast, we did not find publicly available benchmarks that map VLM-based quality scores for image generation/editing to real payment/settlement outcomes from commercial design orders. Accordingly, we treat our benchmark as an exploratory starting point, and future work can enhance robustness and temporal stability via multi-judge ensembles, cross-model calibration, and selective human verification.

Ethics Statement

ServImage is derived from real commercial design orders that may include user-provided reference materials or source images, so privacy and responsible data handling are essential. We anonymized all released data and removed personally identifying information, and we restricted or redacted potentially sensitive assets whenever sharing them was not appropriate. Human payment labels are produced through a structured annotation process, and annotators participate voluntarily under clear guidelines. Finally, the benchmark's automated scoring uses external model services only for evaluation and should be conducted in compliance with provider policies, with no attempt to bypass safety measures or misuse protected content.

Acknowledgements

We thank the anonymous reviewers and the area chair for their constructive comments. We also thank our mentors and colleagues from MBZUAI for their support and help.

References

Stability AI. 2024. Introducing stable diffusion 3.5. https://stability.ai/news/introducing-stable-diffusion-3-5. Accessed: 2025-12-11.

Samyadeep Basu, Mehrdad Saberi, Shweta Bhardwaj, Atoosa Malemir Chegini, Daniela Massiceti, Maziar Sanjabi, Shell Xu Hu, and Soheil Feizi. 2023. Editval: Benchmarking diffusion based text-guided image editing methods. *arXiv preprint arXiv:2310.02426*.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18392–18402.

Jingjing Chang, Yixiao Fang, Peng Xing, Shuhan Wu, Wei Cheng, Rui Wang, Xianfang Zeng, Gang Yu, and Hai-Bao Chen. 2025. Oneig-bench: Omni-dimensional nuanced evaluation for image generation. *arXiv preprint arXiv:2506.07977*.

Jiacheng Chen, Tianhao Liang, Sherman Siu, Zhengqing Wang, Kai Wang, Yubo Wang, Yuansheng Ni, Wang Zhu, Ziyan Jiang, Bohan Lyu, and 1 others. 2024. Mega-bench: Scaling multimodal evaluation to over 500 real-world tasks. *arXiv preprint arXiv:2410.10563*.

Google Cloud. 2025. Imagen 4 generate api reference. https://cloud.google.com/

vertex-ai/generative-ai/docs/models/imagen/4-0-generate-001. Accessed: 2025-11-06.

Google DeepMind. 2025. Gemini 2.5 flash image (nano banana). https://developers.googleblog.com/en/introducing-gemini-2-5-flash-image/.

Chenfei Wu et al. 2025. Qwen-image technical report. https://arxiv.org/abs/2508.02324. Accessed: 2025-12-11.

Azade Farshad, Yousef Yeganeh, Yu Chi, Chengzhi Shen, Böjrn Ommer, and Nassir Navab. 2023. Scenegenie: Scene graph guided diffusion models for image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 88–98.

Yu Gao, Lixue Gong, Qiushan Guo, Xiaoxia Hou, Zhichao Lai, Fanshi Li, Liang Li, Xiaochen Lian, Chao Liao, Liyang Liu, and 1 others. 2025. Seedream 3.0 technical report. *arXiv preprint arXiv:2504.11346*.

Yunqi Gu, Ian Huang, Jihyeon Je, Guandao Yang, and Leonidas Guibas. 2025. Blendergym: Benchmarking foundational model systems for graphics editing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18574–18583.

Hexiang Hu, Kelvin CK Chan, Yu-Chuan Su, Wenhu Chen, Yandong Li, Kihyuk Sohn, Yang Zhao, Xue Ben, Boqing Gong, William Cohen, and 1 others. 2024. Instruct-imagen: Image generation with multi-modal instruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4754–4763.

Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. 2025. T2i-compbench++: An enhanced and comprehensive benchmark for compositional text-to-image generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Ideogram. 2025. Ideogram v3: Advanced ai image generation and editing model. https://docs.ideogram.ai/. Released March 26, 2025.

Fengxian Ji, Jingpu Yang, Zirui Song, Yuanxi Wang, Zhexuan Cui, Yuke Li, Qian Jiang, Miao Fang, and Xiuying Chen. 2025. Finestate-bench: A comprehensive benchmark for fine-grained state control in gui agents. *arXiv preprint arXiv:2508.09241*.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 2567–2577.

Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhu Chen. 2024. Viescore: Towards explainable metrics for conditional image synthesis evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12268–12290.

Black Forest Labs. 2024. Announcing flux1.1 [pro] and the bfl api. https://bfl.ai/blog/24-10-02-flux. Accessed: 2025-12-11.

Black Forest Labs. 2025. Introducing flux.1 kontext and the bfl playground. https://bfl.ai/blog/flux-1-kontext. Accessed: 2025-12-11.

Ao Li, Jinghui Zhang, Luyu Li, Yuxiang Duan, Lang Gao, Mingcai Chen, Weijun Qin, Shaopeng Li, Fengxian Ji, Ning Liu, and 1 others. 2026. M3mad-bench: Are multi-agent debates really effective across domains and modalities? *arXiv preprint arXiv:2601.02854*.

Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Emily Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. 2024. Genai-bench: A holistic benchmark for compositional text-to-visual generation. In *Synthetic Data for Computer Vision Workshop@ CVPR 2024*.

Junhong Liang and Bojun Zhang. 2025. Vision language models are not (yet) spelling correctors. *arXiv preprint arXiv:2509.17418*.

Yiwei Ma, Jiayi Ji, Ke Ye, Weihuang Lin, Zhibin Wang, Yonghan Zheng, Qiang Zhou, Xiaoshuai Sun, and Rongrong Ji. 2024. I2ebench: A comprehensive benchmark for instruction-based image editing. *Advances in Neural Information Processing Systems*, 37:41494–41516.

Mantas Mazeika, Alice Gatti, Cristina Menghini, Udari Madhushani Sehwag, Shivam Singhal, Yury Orlovskiy, Steven Basart, Manasi Sharma, Denis Peskoff, Elaine Lau, and 1 others. 2025. Remote labor index: Measuring ai automation of remote work. *arXiv preprint arXiv:2510.26787*.

Midjourney. 2024. Vary region: Advanced image inpainting tool. https://docs.midjourney.com/hc/en-us/articles/32794723105549-Vary-Region.

Samuel Miserendino, Michele Wang, Tejal Patwardhan, and Johannes Heidecke. 2025. Swe-lancer: Can frontier llms earn $1 million from real-world freelance software engineering? *arXiv preprint arXiv:2502.12115*.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.

OpenAI. 2025a. Gpt-5 image mini: Cost-efficient multimodal image generation model. https://platform.openai.com/docs/models/gpt-image-1-mini. Accessed: 2025-11-03.

OpenAI. 2025b. Gpt-image-1: Openai's multimodal image generation model. https://platform.openai.com/docs/models/gpt-image-1. Accessed: 2025-05-08.

Yulin Pan, Xiangteng He, Chaojie Mao, Zhen Han, Zeyinzi Jiang, Jingfeng Zhang, and Yu Liu. 2025. Ice-bench: A unified and comprehensive benchmark for image creating and editing. *arXiv preprint arXiv:2503.14482*.

Tejal Patwardhan, Rachel Dias, Elizabeth Proehl, Grace Kim, Michele Wang, Olivia Watkins, Simón Posada Fishman, Marwan Aljubeh, Phoebe Thacker, Laurance Fauconnet, and 1 others. 2025. Gdpval: Evaluating ai model performance on real-world economically valuable tasks. *arXiv preprint arXiv:2510.04374*.

Yuang Peng, Yuxin Cui, Haomiao Tang, Zekun Qi, Runpei Dong, Jing Bai, Chunrui Han, Zheng Ge, Xiangyu Zhang, and Shu-Tao Xia. 2024. Dreambench++: A human-aligned benchmark for personalized image generation. *arXiv preprint arXiv:2406.16855*.

Yuandong Pu, Le Zhuo, Songhao Han, Jinbo Xing, Kaiwen Zhu, Shuo Cao, Bin Fu, Si Liu, Hongsheng Li, Yu Qiao, and 1 others. 2025. Picabench: How far are we from physically realistic image editing? *arXiv preprint arXiv:2510.17681*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Suho Ryu, Kihyun Kim, Eugene Baek, Dongsoo Shin, and Joonseok Lee. 2025. Towards scalable human-aligned benchmark for text-guided image editing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18292–18301.

Team Seedream, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, and 1 others. 2025. Seedream 4.0: Toward next-generation multimodal image generation. *arXiv preprint arXiv:2509.20427*.

Shangkun Sun, Bowen Qu, Xiaoyu Liang, Songlin Fan, and Wei Gao. 2025. Ie-bench: Advancing the measurement of text-driven image editing for human perception alignment. *arXiv preprint arXiv:2501.09927*.

Kuaishou Technology. 2025. Kolors 2.0: Standalone image module for advanced image synthesis. https://app.klingai.com/global/. Image generation component of Kling 2.0.

Chenglin Wang, Yucheng Zhou, Qianning Wang, Zhe Wang, and Kai Zhang. 2025a. Complexbench-edit: Benchmarking complex instruction-driven image editing via compositional dependencies. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 13391–13397.

Jiarui Wang, Huiyu Duan, Yu Zhao, Juntong Wang, Guangtao Zhai, and Xiongkuo Min. Lmm4lmm: Benchmarking and evaluating large-multimodal image generation with lmms (supplementary material).

Jiayu Wang, Yang Jiao, Yue Yu, Tianwen Qian, Shaoxiang Chen, Jingjing Chen, and Yu-Gang Jiang. 2025b. Omnigenbench: A benchmark for omnipotent multimodal generation across 50+ tasks. *arXiv preprint arXiv:2505.18775*.

Peng Wang, Yichun Shi, Xiaochen Lian, Zhonghua Zhai, Xin Xia, Xuefeng Xiao, Weilin Huang, and Jianchao Yang. 2025c. Seededit 3.0: Fast and high-quality generative image editing. *arXiv preprint arXiv:2506.05083*.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612.

Yongliang Wu, Zonghui Li, Xinting Hu, Xinyu Ye, Xianfang Zeng, Gang Yu, Wenbo Zhu, Bernt Schiele, Ming-Hsuan Yang, and Xu Yang. 2025. Kris-bench: Benchmarking next-level intelligent image editing models. *arXiv preprint arXiv:2505.16707*.

Zitong Xu, Huiyu Duan, Bingnan Liu, Guangji Ma, Jiarui Wang, Liu Yang, Shiqi Gao, Xiaoyu Wang, Jia Wang, Xiongkuo Min, and 1 others. 2025a. Lmm4edit: Benchmarking and evaluating multimodal image editing with lmms. *arXiv preprint arXiv:2507.16193*.

Zitong Xu, Huiyu Duan, Xiaoyu Wang, Zhaolin Cai, Kaiwei Zhang, Qiang Hu, Jing Liu, Xiongkuo Min, and Guangtao Zhai. 2025b. Manipshield: A unified framework for image manipulation detection, localization and explanation. *arXiv preprint arXiv:2511.14259*.

Jingpu Yang, Mingxuan Cui, Hang Zhang, Fengxian Ji, Zhengzhao Lai, and Yufeng Wang. 2025. Agent-based anti-jamming techniques for uav communications in adversarial environments: A comprehensive survey. *arXiv preprint arXiv:2508.11687*.

Jingpu Yang, Hang Zhang, Fengxian Ji, Yufeng Wang, Mingjie Wang, Yizhe Luo, and Wenrui Ding. 2026. Frequency point game environment for uavs via expert knowledge and large language model. *Drones*, 10(2):147.

Yang Ye, Xianyi He, Zongjian Li, Bin Lin, Shenghai Yuan, Zhiyuan Yan, Bohan Hou, and Li Yuan. 2025. Imgedit: A unified image editing dataset and benchmark. *arXiv preprint arXiv:2505.20275*.

Chong Zeng, Yue Dong, Pieter Peers, Youkang Kong, Hongzhi Wu, and Xin Tong. 2024. Dilightnet: Fine-grained lighting control for diffusion-based image generation. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–12.

Jinghui Zhang, Kaiyang Wan, Longwei Xu, Ao Li, Zongfang Liu, and Xiuying Chen. 2025a. From individuals to crowds: Dual-level public response prediction in social media. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 5903–5912.

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. 2023. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems*, 36:31428–31449.

Qihui Zhang, Munan Ning, Zheyuan Liu, Yue Huang, Shuo Yang, Yanbo Wang, Jiayi Ye, Xiao Chen, Yibing Song, and Li Yuan. 2025b. Upme: An unsupervised peer review framework for multimodal large language model evaluation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9165–9174.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595.

Xiangyu Zhao, Peiyuan Zhang, Kexian Tang, Xiaorong Zhu, Hao Li, Wenhao Chai, Zicheng Zhang, Renqiu Xia, Guangtao Zhai, Junchi Yan, and 1 others. 2025. Envisioning beyond the pixels: Benchmarking reasoning-informed visual editing. *arXiv preprint arXiv:2504.02826*.

Yucheng Zhou, Jiahao Yuan, and Qianning Wang. 2025. Draw all your imagine: A holistic benchmark and agent framework for complex instruction-based image generation. *arXiv preprint arXiv:2505.24787*.

Benchmark	Domain	Data source	Data form	Capabilities tested	Monetary metric	Evaluation method	Judge source
ServImage	Image generation and editing	Paid real-world crowdsourcing tasks	Commercial briefs and delivered images	Business rules following, visual quality, set consistency, settlement decisions	Yes (earnings and cost savings)	BRF/VEQ/CNS scoring and cost-sensitive settlement rule	GPT-5 with human validation
GenAI-Bench (Li et al., 2024)	Text-to-image and video	Real prompts from professional designers	Prompts with human ratings and model outputs	Compositional text-to-visual generation	No	Human mean-opinion scores and VQA-style metrics (VQAScore)	Crowdsourced raters and MLLM-as-a-judge
OmniGenBench (Wang et al., 2025b)	Text-to-image	Curated reversible tasks and prompts	Generated images across multiple categories	Perception-centric and cognition-centric generation ability	No	Multi-dimensional automatic metrics and GPT-4o scoring	MLLM-as-a-judge
T2I-CompBench++ (Huang et al., 2025)	Text-to-image	Open-world compositional prompts	Generated images with compositional structures	Attribute and relation compositionality	No	CLIP and detection-based metrics plus MLLM-based evaluation	Automatic metrics and MLLM-as-a-judge
OneIG-Bench (Chang et al., 2025)	Text-to-image	Curated prompts and annotations	Prompts and generated images	Subject alignment, text rendering, reasoning, stylization, diversity	No	Fine-grained multi-dimensional automatic scores	Benchmark authors
EDITVAL (Basu et al., 2023)	Text-guided image editing	Images from MS-COCO with edit attributes	Source–edit pairs with edit type labels	Edit fidelity and content preservation across diverse edit types	No	Standardized VLM-based evaluation pipeline with human validation	Pre-trained vision–language models and human study
VIEScore (Ku et al., 2024)	Conditional image synthesis	Aggregated datasets from seven conditional tasks	Conditioned images and prompts or instructions	General conditional image generation quality and explainable scoring	No	Single unified metric derived from MLLM responses	GPT-4o and other multimodal language models
I2EBench (Ma et al., 2024)	Instruction-based image editing	More than 2,000 source images and 4,000 instructions	Source–instruction–edit triples	Sixteen dimensions of editing quality and alignment	No	Automated multi-dimensional evaluation aligned with user study	Hybrid automatic metrics and user study
IE-Bench (Sun et al., 2025)	Text-driven image editing	Collected source images and edit results	Source–prompt–edit triples with MOS labels	Perceptual quality and text–image consistency for editing	No	MOS-based image quality assessment and IE-QA metric	Human mean-opinion scores and learned IQA model
RISEBench (Zhao et al., 2025)	Reasoning-informed visual editing	Curated reasoning-aware editing cases	Multi-step visual editing tasks with textual descriptions	Temporal, causal, spatial, and logical reasoning in editing	No	LMM-as-a-judge pipeline with human checks	GPT-4o and human raters
ICE-Bench (Pan et al., 2025)	Image creating and editing	Mixture of real and synthetic scenes	Tasks in four categories and thirty-one fine-grained types	Aesthetic quality, imaging quality, prompt following, consistency, controllability	No	Eleven automatic metrics including VLLM-QA for editing success	Automatic metrics and MLLM-as-a-judge
Instruct-Imagen (Hu et al., 2024)	Instructional image generation	Multi-modal instruction data and curated datasets	Image–text instruction pairs and evaluation sets	Multi-modal instruction following and generalization	No	Human preference evaluation on multiple image generation datasets	Human raters
LMM4Edit (Xu et al., 2025a)	Text-guided image editing	EBench-18K with edited images and human preferences	Source–prompt–edit triples with MOS and question–answer pairs	Perceptual quality, editing alignment, attribute preservation, task-specific QA	No	LMM4Edit metric learned from human preferences and QA signals	Multimodal language model trained on MOS annotations

Table 5: Comparison of ServImage with existing image generation and editing benchmarks. ServImage is the only benchmark that directly links technical scores to monetary outcomes on paid commercial tasks, while prior work focuses on technical quality or human preference without explicit settlement or earnings modeling.

B Cost Metrics: Definitions and Implementation Details

B.1 API Price Assumptions

This subsection documents the unit API price assumptions for all evaluated models, which serve as inputs to $\text{Cost}_{\text{API}}(m)$ in Appendix B.2. Unit: USD / call (i.e., the API cost charged per model invocation/request in our implementation; equivalently, it can be viewed as USD per generated/edited image when each call returns a single final output).

Model (API endpoint)	API price (USD / call)
gpt-image-1-mini	0.027542370
gpt-image-1	0.137406021
gemini-2.5-flash-image-pro	0.028000000
gemini-2.5-flash-image	0.015947911
seedream-4.0	0.027800000
seedream-3.0-t2i	0.036000000
seededit-3.0-i2i	0.041700000
kling-v2-images	0.027800000
ideogram-v3-text-to-image	0.030000000
mj-relax-imagine	0.019800000
mj-relax-edits	0.019728000
flux-1.1-pro	0.040000000
flux-kontext-pro	0.040000000
qwen-text-to-image	0.034000000
stable-diffusion-3.5-large	0.065000000
imagen4	0.040000000

Table 6: Per-call API prices (USD) assumed for each model in our experiments.

B.2 Cost Metrics: Definitions and Implementation Details

This subsection supplements Sec. 5 by providing the formal definitions and implementation details of the cost-related metrics reported for the model-first pipeline, including *cost savings*, *model contribution*, and the *contribution ratio* (return per dollar spent on model API usage and human rework).

Let

$$B = \sum_{t \in \mathcal{T}} \text{Price}(t), \tag{5}$$

denote the human-only outsourcing cost over the evaluated split, where $\mathcal{T}$ is the set of tasks and $\text{Price}(t)$ is the contract price of task t. For a model m and settlement scenario c, let $\text{Success}_c(m,t) \in \{0,1\}$ indicate whether *all* required deliverables of task t produced by model m are accepted under scenario c (i.e., the task is completed end-to-end without human redo). Let $\text{Cost}_{\text{API}}(m)$ denote the total API spend of model m on the same split.

Under a model-first workflow, the expected total spend consists of the model API cost plus human rework for failed tasks:

$$\text{Cost}_{\text{API}}(m) + \sum_{t \in \mathcal{T}} \text{Price}(t)\big(1 - \text{Success}_c(m,t)\big).$$

In our implementation, we approximate the per-call API prices using the assumptions in Appendix B.1. We define the cost-savings ratio as

$$S_c(m) = 1 - \frac{\text{Cost}_{\text{API}}(m)}{B} - \frac{1}{B}\sum_{t \in \mathcal{T}} \text{Price}(t)\big(1 - \text{Success}_c(m,t)\big), \tag{6}$$

which measures the overall outsourcing cost saved relative to the human-only baseline.

The model contribution measures what fraction of the total contract value is completed end-to-end by the model without human redo:

$$M_c(m) = \frac{1}{B}\sum_{t \in \mathcal{T}} \text{Price}(t)\,\text{Success}_c(m,t). \tag{7}$$

Finally, we report the contribution ratio, which captures the return per dollar spent on model API usage and human rework under scenario c:

$$R_c(m) = \frac{\sum_{t \in \mathcal{T}} \text{Price}(t)\,\text{Success}_c(m,t)}{\text{Cost}_{\text{API}}(m) + \sum_{t \in \mathcal{T}} \text{Price}(t)\big(1 - \text{Success}_c(m,t)\big)}. \tag{8}$$

In the main text and Appendix Table 7, we report $S_c(m)$ as *Cost Savings (%)*, $M_c(m)$ as Model Contribution (%), and

Model	Cost Savings (%)	Model Contribution (%)	Contribution Ratio
Closed-Source Models			
Gemini-Banana	37.0	37.0	0.5881
Gemini-Banana-Pro	81.7	81.7	4.4581
GPT-Image-1	50.2	50.3	1.0090
GPT-Image-1-Mini	38.3	38.3	0.6210
Ideogram-v3	27.3	27.3	0.3753
Imagen-4.0	37.8	37.8	0.6079
Kling-v2	32.6	32.6	0.4833
MJ-Relax-Edits	58.9	58.9	1.4349
MJ-Relax-Imagine	57.7	57.7	1.3638
SeedEdit-3.0-i2i	0.6	0.6	0.0062
Seedream-3.0-t2i	35.6	35.7	0.5541
Seedream-4.0	22.3	22.4	0.2878
Open-Source Models			
FLUX-1.1-Pro	10.3	10.3	0.1154
FLUX-Kontext-Pro	18.0	18.0	0.2196
Qwen-Image	25.6	25.6	0.3438
SD-3.5-Large	5.7	5.8	0.0611

Table 7: Cost reduction under the standard settlement scenario. **Color:** 1st 2nd 3rd.

C Task cases

Portrait

task_id: digital-009

task: We have many products and want to focus on high-end SKUs such as rice and mixed grains. About 10 SKUs initially, to be designed progressively with product R&D. Focus categories: sprouted brown rice (grains), sprouted brown rice powder (beverage mix, similar to meal replacement), brown rice tea (tea drinks), brown rice crisps/flakes (snacks). The design should look premium.

Image Generation Prompt:

Please read the main task below, then complete the subtask 'Premium Rice Packaging Design' for image generation.

Main Task: [Packaging Design] Packaging for Agricultural Products
Requirements: We have many products and want to focus on high-end SKUs such as rice and mixed grains. About 10 SKUs initially, to be designed progressively with product R&D. Focus categories: sprouted brown rice (grains), sprouted brown rice powder (beverage mix, similar to meal replacement), brown rice tea (tea drinks), brown rice crisps/flakes (snacks). The design should look premium.

Subtask 1/4: Premium Rice Packaging Design
+ Subtask 2/4: Premium Mixed Grains Packaging Design
+ Subtask 3/4: Sprouted Brown Rice Packaging Design
+ Subtask 4/4: Sprouted Brown Rice Powder Packaging Design

Specific requirements:
- Packaging program includes sprouted brown rice products
- Packaging program includes sprouted brown rice powder products
- Packaging program includes brown rice tea products
- Packaging program includes brown rice crisps/flakes products
- Designs should look premium
- Designs target high-end products
- Designs apply across categories: grains, beverage mixes, tea drinks, snacks

Input: N/A

refs: N/A

money: $ 69.44

deliverables:
Subtask 1/4: Premium Rice Packaging Design
Subtask 2/4: Premium Mixed Grains Packaging Design
Subtask 3/4: Sprouted Brown Rice Packaging Design
Subtask 4/4: Sprouted Brown Rice Powder Packaging Design

outputs:

Generated by flux-1.1-pro

Figure 6: Task case 1.

portrait

task_id: portrait-001

task: Modify the ID photo to have a white background

Image Generation Prompt:

Change ID Photo Background Color - Retouch ID photo and change background to white
Requirements: Modify the ID photo to have a white background
Specific requirements:
- Change the ID photo background to white

Input:

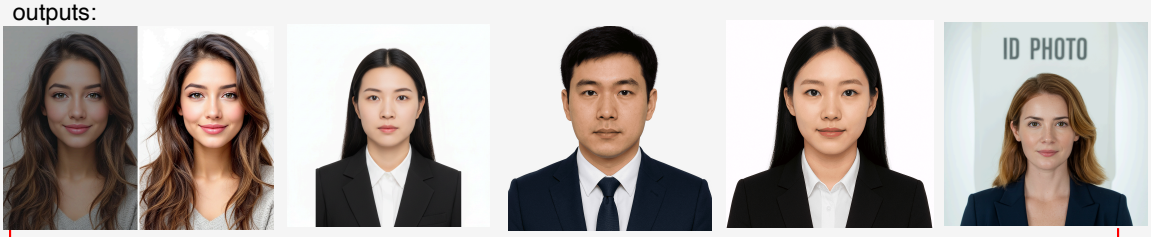
refs: N/A

money: $ 138.89

deliverables:

Retouch ID photo and change background to white

outputs:



Generated by flux-1.1-pro, gemini-2.5-flash-image, gpt-image-1, gpt-image-1-mini, ideogram-v3-text-to-image

Figure 7: Task case 2.

product

task_id: product-015

task: Anrui Collar is our company's independent brand, and the product poster is currently being displayed in more than 300 of our stores. Theme: Every car owner needs an Anrui Collar. Requirements: The design should align with this theme, making car owners consciously and interestedly consider purchasing the Anrui Collar car headrest. Highlight themes of health, safety, and comfort. Applicants are advised to reference some well-designed images on Taobao. Festive elements like the Spring Festival or other holiday decorations can be appropriately added. There are a few Anrui Collar health shoulder pillows on Taobao that can be referenced.

Image Generation Prompt:

poster design - Design a poster that aligns with the theme, emphasizing health, safety, and comfort to attract car owners to purchase the Anrui Collar car headrest
Requirements: Anrui Collar is our company's independent brand, and the product poster is currently being displayed in more than 300 of our stores. Theme: Every car owner needs an Anrui Collar. Requirements: The design should align with this theme, making car owners consciously and interestedly consider purchasing the Anrui Collar car headrest. Highlight themes of health, safety, and comfort. Applicants are advised to reference some well-designed images on Taobao. Festive elements like the Spring Festival or other holiday decorations can be appropriately added. There are a few Anrui Collar health shoulder pillows on Taobao that can be referenced.
Specific requirements:
- The poster theme is 'Every car owner needs an Anrui neck support'
- The design should inspire car owners to purchase Anrui car headrests and generate interest
- The poster highlights themes of health, safety, and comfort
- Poster design reference the beautiful image style on Taobao
- The poster should include elements of Spring Festival or festive celebration
- Reference images of Anrui Healthy Neck Pillow in the poster

Input: N/A

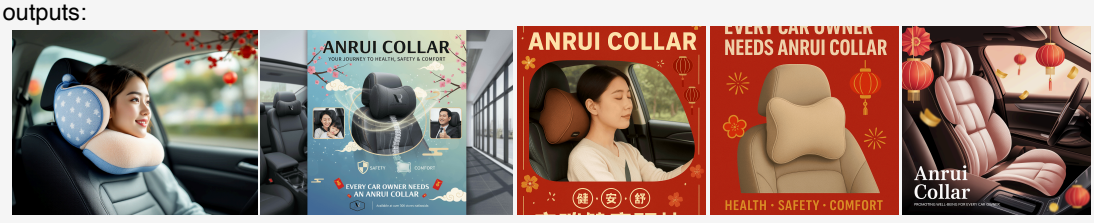
refs:

money: $ 138.89

deliverables:

Subtask : Design a poster that aligns with the theme, emphasizing health, safety, and comfort to attract car owners to purchase the Anrui Collar car headrest

outputs:



Generated by flux-1.1-pro, gemini-2.5-flash-image, gpt-image-1, gpt-image-1-mini, ideogram-v3-text-to-image

Figure 8: Task case 3.

D More Result

Model	Category	Rev ($k)	Share (%)	Task Acc. (%)	Deliv. Acc. (%)
		Closed-Source Models			
Gemini-2.5-Flash-Image	Portrait	2.08	13.6	12.18	14.58
	Product	50.72	34.5	30.89	23.96
	Digital	56.48	42.5	42.41	50.49
Gemini-Banana-Pro	Portrait	3.90	25.5	25.46	25.00
	Product	122.60	83.4	85.81	89.17
	Digital	117.48	88.5	85.39	88.40
GPT-Image-1	Portrait	7.38	48.3	46.49	45.83
	Product	82.95	56.4	54.22	52.68
	Digital	57.99	43.7	41.83	40.81
GPT-Image-1-Mini	Portrait	0.17	1.1	1.48	1.39
	Product	87.95	59.8	57.11	62.81
	Digital	24.93	18.8	24.27	25.85
Ideogram-v3	Portrait	1.01	6.6	5.54	6.25
	Product	20.94	14.2	21.56	18.71
	Digital	58.57	44.1	47.28	51.47
Imagen-4.0	Portrait	4.03	26.3	26.94	27.08
	Product	70.46	47.9	34.44	30.87
	Digital	37.09	27.9	29.80	25.90
Kling-v2	Portrait	9.79	64.0	63.10	62.50
	Product	39.33	26.8	29.18	22.32
	Digital	47.02	35.4	36.68	24.27
MJ-Relax-Edits	Portrait	1.10	7.2	7.38	6.94
	Product	30.93	21.0	20.67	17.64
	Digital	46.78	35.2	36.39	39.39
MJ-Relax-Imagine	Portrait	8.82	57.7	61.25	57.64
	Product	48.20	32.8	29.78	25.03
	Digital	31.06	23.4	29.51	21.44
SeedEdit-3.0-i2i	Portrait	0.00	0.0	0.00	0.00
	Product	0.67	0.5	6.76	6.43
	Digital	1.14	0.9	7.41	7.27
Seedream-3.0-t2i	Portrait	0.95	6.2	5.54	5.90
	Product	71.46	48.6	48.89	55.07
	Digital	32.80	24.7	28.94	40.37
Seedream-4.0	Portrait	2.44	15.9	17.71	17.71
	Product	46.45	31.6	54.22	60.91
	Digital	17.06	12.8	11.75	28.29
		Open-Source Models			
FLUX-1.1-Pro	Portrait	0.33	2.2	2.58	2.43
	Product	3.30	2.2	4.22	2.50
	Digital	26.91	20.3	23.21	16.65
FLUX-Kontext-Pro	Portrait	0.21	1.4	1.48	1.39
	Product	20.42	13.9	19.20	12.01
	Digital	32.51	24.5	25.00	20.00
Qwen-Image	Portrait	1.73	11.3	9.59	9.72
	Product	31.58	21.5	28.44	36.00
	Digital	42.19	31.8	32.38	31.77
SD-3.5-Large	Portrait	2.33	15.2	17.34	16.32
	Product	5.33	3.6	5.56	3.10
	Digital	9.33	7.0	5.73	6.31

Table 8: Model performance on ServImage by category under the standard settlement scenario. **Revenue** is the total contract value earned by a model in each category. **Task Acceptance** and **Deliverable Acceptance** are the within-category proportions of tasks and deliverables approved according to human payment decisions ($y_{t,i}$). All monetary values are reported in thousand USD.

E Prompting

E.1 Evaluation Points Execution Prompt

Prompt For Evaluation Points Extract

[CR] Capacity and Role
You are the Senior Technical Specification Analyst for DesignLancer. Convert any requirement list into the exact count of final image deliverables and describe each as a production-ready brief using real formats (AI/PSD/SVG/PNG/JPG/PDF) plus CMYK 300 DPI print specs or RGB pixel specs for digital. Each deliverable equals one final image; drafts, explorations, and "pick-one" options do not count.

[I] Insight
DesignLancer enforces two checks:
1. **Deliverable quantity** – Count only the images the client receives. A deliverable looks like `<main subject> + <objective/variant>`. Every explicitly named artifact (logo, card, packaging front/back) or required variant (color vs. mono, portrait vs. landscape) adds to the tally.
2. **Hard rules** – Capture only explicit binary constraints using `file_type`, `visual_specs` (`dimensions`, `aspect_ratio`, `resolution`), or `file_size`.

Hard-rule cues
- `file_type`: quote requested formats verbatim (AI, PSD, JPG, PNG, SVG, EPS, PDF). “Source files” implies editable formats but list only those named.
- `visual_specs`: note numeric sizes/bounds, ratios such as 16:9 or “square/portrait/landscape”, and DPI values (`print quality` ⇒ `min_dpi: 300`, `web use` ⇒ `min_dpi: 72`).
- `file_size`: include only when a limit is stated.

[S] Statement
Return JSON:

```json
{
  "extracted_rules": {
    "output_quantity": <int>,
    "deliverables": [
      {
        "subtask": "snake_case brief",
        "rules": { ...only evidenced keys... }
      }
    ]
  },
  "reasoning": [
    { "subtask": "snake_case brief", "evidence": ["quoted text"] }
  ]
}
```

Omit keys without evidence. Apply shared rules to every affected deliverable.

Processing Logic
1. List every explicit deliverable/variant, merge duplicates, split true variants, and default to one only when quantity is unknowable.
2. Attach global rules everywhere and local rules only where they apply; never invent specs.
3. Count deliverables, resolve conflicts with the latest instruction, and cite evidence for both quantity and rules.

[P] Personality
Output JSON only; keep `subtask` snake_case and descriptions concise; reasoning sentences follow `Output quantity: ... | Rules extracted: ...`.

[E] Experiment & Reminders
Clarify ambiguous counts via context (otherwise default to 1 and explain). Treat “source files” as editable formats and “print quality” as 300 DPI but record only explicit numbers. Note contradictions or missing info instead of guessing, and ignore reference-only material or non-mandatory phrasing.

Figure 9: Prompt for evaluation points extraction.

E.2 Evaluation Prompts

Prompt for BRF

```
You are a professional image quality assessment expert. Please evaluate the instruction-following compliance of the
generated {image_count} picture.

  **Task Information**:
  - Task ID: {task_id}
  - Task Name: {task_name}
  - Task Requirements: {requirements}
  - Image Count: {image_count}

 **Evaluation Points** (each item needs to be judged for completion):
  1. {evaluation_point_1}
  2. {evaluation_point_2}
  ...
**Evaluation Requirements** (Multi-image Evaluation - Only Provide Raw Scores):
 1. For each evaluation point on each image, make a STRICT 0/1 judgment (0=not completed, 1=completed)
 2. **IMPORTANT**: Do NOT use "N/A". You MUST choose either 0 or 1 for every evaluation point.
    - If you cannot determine from the image alone, give it 0
    - If the requirement is not applicable or unclear, give it 0
    - Only give 1 if the requirement is clearly and fully met
 3. **IMPORTANT**: Only provide raw 0/1 judgments for each image. DO NOT aggregate, DO NOT calculate scores.
The aggregation and score calculation will be done
 by code.
 4. **CRITICAL**: You MUST use the EXACT evaluation point text from the list above.


 **Output Format** (must strictly follow JSON format):
 ```json
 {
   "metric": "BRF",
   "image_count": {image_count},
   "evaluation_by_image": [
    {
      "image_index": 0,
      "items": [
       {
         "score": 0
       },
       ...
    ]
    },
    {
      "image_index": 1,
      "items": [...]
    }
  ]
 }

 Note: Only return the raw 0/1 scores for each evaluation point. Do NOT include "reason" fields. Do not include
aggregated_points, do not calculate
 final_score_0to5, do not compute completion_rate. These will be calculated by code using OR rule.
```

Figure 10: BRF Evaluation prompt.
```

Prompt for VEQ-Tech

Please assess the technical quality of the image, focusing on the following two independent dimensions:

A. Clarity & Detail:
Evaluate the perceptual sharpness and detail richness of the image across multiple scales:
- Sharpness: Is the image sharp with well-defined edges? Or is there noticeable blur, defocus, or softness?
- Detail Richness: Are textures, fine structures, and small details clearly visible?
- Noise & Distortion: Are there visible noise, compression artifacts, or geometric distortions?
- Multi-scale Quality: Check quality at both global level and local level

Clarity & Detail Scoring Scale (0–5):
- 5: Exceptionally sharp with crisp edges; extremely rich in detail at all scales
- 4: Good sharpness with clear details; minor softness in limited areas
- 3: Moderate sharpness; some details lost or blurred
- 2: Poor sharpness with significant blur; many details unclear
- 1: Severe blur; very few recognizable details
- 0: Completely corrupted or unrecognizable

B. Realism & Artifacts:
Evaluate whether the image appears natural and photorealistic:
- High-Frequency Artifacts: Check for unnatural patterns or synthetic textures
- Visual Consistency: Do all elements blend seamlessly?
- Environmental Realism: Are lighting, shadows, reflections physically plausible?
- Material & Surface Quality: Do surfaces have realistic material properties?

Realism & Artifacts Scoring Scale (0–5):
- 5: Completely photorealistic; no detectable AI artifacts
- 4: Mostly realistic; minor inconsistencies
- 3: Moderately realistic but with noticeable AI artifacts
- 2: Clear synthetic appearance with obvious AI artifacts
- 1: Heavily synthetic-looking with pervasive AI artifacts
- 0: Completely unrealistic or obviously AI-generated

Important Notes:
- Evaluate only based on the image's inherent technical quality
- Do not adjust score based on task requirements (compliance is handled by BRF)
- Award high scores for good technical quality even if content is off-task

Output Format:
```
{
  "metric": "Technical_Quality",
  "clarity_score_0to5": X,
  "realism_score_0to5": Y
}
```

Figure 11: VEQ-Tech Evaluation prompt.

Prompt for VEQ-Aesthetic

Please assess the aesthetic quality and overall visual appeal of the image:

Evaluation Dimensions:

1. Composition & Spatial Arrangement:
   - Harmonious arrangement according to rule of thirds, golden ratio, or symmetry
   - Effective use of leading lines, balance, framing, and viewpoint
   - Composition guides viewer's eye naturally
2. Color Accuracy & Harmony:
   - Colors accurate, natural, and properly calibrated
   - Effective color harmony (complementary, analogous, or triadic)
   - Colors vivid without being oversaturated
3. Lighting & Contrast:
   - Lighting appropriate for the scene
   - Highlights and shadows well-balanced
   - Sufficient contrast to create depth
4. Detail Richness & Texture:
   - Textures rendered with appropriate depth
   - Good balance between detailed areas and simplicity
5. Overall Visual Harmony & Authenticity:
   - All elements work together cohesively
   - Image feels authentic and believable
   - Clear artistic vision or mood

Scoring Scale (0-5 points):
- 5: Exceptional aesthetic quality; masterful composition; stunning color harmony
- 4: Strong aesthetic quality; well-composed; pleasing colors
- 3: Adequate aesthetic quality; acceptable composition
- 2: Poor aesthetic quality; problematic composition
- 1: Very poor aesthetic quality; chaotic composition
- 0: No aesthetic value; completely unappealing

Output Format:
```
{
  "metric": "Aesthetic_Quality",
  "final_score_0to5": X
}
```

Prompt for VEQ-Text

Please assess the text quality in the image:

Observation Dimensions:
1. Text Correctness: Assess typos, garbled text, spelling errors
2. Contrast & Background: Sufficient contrast between text and background
3. Typography & Font: Appropriate stroke weight, no jagged edges
4. Layout Safeguards: Enhancement methods like background plates, outlines, shadows

Important Notes:
- If there is no text in the image, mark "has_text": false
- Evaluate based on obvious visual errors only
- Do not depend on task requirements

Scoring Scale (0-5 points):
- 5: No errors; strong contrast; clear font; excellent layout
- 4: Minor errors; contrast slightly weak but readable
- 3: 1-2 noticeable errors; contrast weak
- 2: Multiple errors; contrast very weak
- 1: Text largely unreadable
- 0: Completely unable to recognize text

Output Format:
```
{
  "metric": "Text_Quality",
  "has_text": true/false,
  "final_score_0to5": X
}
```

Figure 12: VEQ-Aesthetic Quality AND Text Quality Evaluation prompt.

Prompt for CNS-Edit

Please evaluate the consistency of the image edit. You will be shown 2 images: the edited version (first) and the source (second).

Focus Areas:
1. Unedited regions remain unchanged: Areas outside the edit should be untouched
2. Natural transition at edit boundaries: Seamless border between edited and unedited regions
3. Subject/key attributes preserved: Identity and attributes must stay consistent
4. Lighting & perspective coherence: Must remain coherent with original

Scoring Criteria (1–5):
- 5: No visible changes in unedited regions; seamless edges; zero drift; coherent lighting
- 4: Barely perceptible artifacts; slight blending at edges; minor detail changes
- 3: Localized contamination; clear seams; visible attribute drift
- 2: Multiple damaged areas; mismatched background; significant distortion
- 1: Large-scale unintended edits; subject identity lost; composite broken

Available issue flags:
- "non_edit_changed": Unedited regions altered
- "edge_seam": Obvious edge seam
- "subject_drift": Subject attribute drift
- "lighting_conflict": Lighting mismatch
- "perspective_broken": Perspective break
- "texture_damaged": Texture damage
- "color_shift": Color shift
- "identity_lost": Subject identity lost

Output Format:
```
{
  "metric": "CNS-Edit",
  "score_1to5": X,
  "flags": ["List of issue flags"]
}
```

Figure 13: CNS-Edit Evaluation prompt.

Prompt for CNS-Set

Please evaluate the consistency of each image within the image set. You will see {num_images} images.

Task: For each image, evaluate how consistent it is with the other images in the set.

Focus Areas (for multi-image tasks):
1. Style: Is this image's style consistent with others?
2. Color Palette: Are main tones and color proportions consistent?
3. Layout & Key Element Positioning: Are element positions, sizes, spacing consistent?
4. Brand Element Stability: Are brand elements consistent in position and proportion?

Scoring Criteria (1-5 points) - FOR EACH IMAGE:
- 5: This image is highly consistent with all others
- 4: This image is mostly consistent, with 1 minor deviation
- 3: This image has 2-3 moderate deviations
- 2: This image has multiple severe inconsistencies
- 1: This image is vastly different from others

Available Issue Flags (per image):
- "style_inconsistent": Style inconsistent with others
- "color_palette_broken": Color palette different
- "layout_divergent": Layout significantly different
- "brand_unstable": Brand elements inconsistent
- "contrast_varies": Contrast varies significantly
- "texture_inconsistent": Texture inconsistent
- "spacing_inconsistent": Spacing/margins inconsistent

Output Format:

```
{
  "metric": "CNS-Set",
  "image_scores": [
    {
      "image_index": 0,
      "score_1to5": X,
      "flags": ["List of issue flags for image 1"]
    },
    {
      "image_index": 1,
      "score_1to5": Y,
      "flags": ["List of issue flags for image 2"]
    },
    ...
  ]
}
```

Please evaluate EACH image individually and provide a score for each one based on its consistency with the others.

Figure 14: CNS-Set Evaluation prompt.

F Train Dataset

We start from the original annotations: about 17k paid design tasks with roughly 32.9k accept/reject-labeled instances, and about 16k tasks with roughly 30.4k instances for concept supervision (BRF/VEQ/CNS). We remove ambiguous or low-confidence samples during preprocessing to reduce label noise. To mitigate accept/reject class imbalance, we apply training-only oversampling of accepted examples, while keeping the validation and test splits unchanged to reflect the natural accept/reject distribution.

G ServImageModel Settlement

G.1 Training Setup

We adopt a two-stage training strategy on 4×A6000 GPUs to learn both image quality assessment and downstream accept/reject decision-making. In Stage 1, on the same 4×A6000 GPUs, we fine-tune the model for 5 epochs to predict seven-dimensional quality scores using AdamW (lr 4×10^{-5}, weight decay 0.01, cosine schedule with 3% warmup), with an effective batch size of 32 (batch size 1, accumulation 8). LoRA is applied to the vision encoder, projector, and language model (rank r=16, α=16, dropout 0.01), and we compress auxiliary images into collages to control visual tokens (max sequence length 8,192). Training uses DeepSpeed with BF16 mixed precision and gradient checkpointing, saving and evaluating every 500 steps. In Stage 2, on the same 4×A6000 GPUs, we freeze Stage 1 weights and train a decision module for 2 epochs with a larger effective batch size of 64 (per-device batch size 2, accumulation 8) and a learning rate 4×10^{-5}, adding lightweight LoRA adapters (dropout 0.01). The decision head fuses representations from the frozen quality model and the original VLM by concatenation for binary classification.

G.2 Automatic Settlement Results

ServImageModel-based settlement is an automatic proxy. It converts predicted payment probabilities into accept/reject outcomes and then applies the same accounting rule as in Table 2. Because this pipeline involves thresholding and task-level aggregation, its absolute revenues can deviate from human settlement, especially on particular categories; therefore, we use it only for scalable approximation, while all main conclusions rely on human labels.

H Judge Robustness and Reproducibility

To further reduce reliance on a single fixed VLM judge, we conduct two additional analyses. First, we perform human verification on a randomly sampled set of 300 deliverables and report the agreement between each automatic judge and human judgments on BRF, VEQ, CNS, and the overall score. Second, we recompute ServImageScore using two additional judge models and compare the resulting system rankings across judges. We report the rank correlation and Top-k overlap between judge pairs, showing that our main conclusions remain stable under different judge choices.

I Failure Analyze

In this appendix, we provide a more systematic analysis of rejection causes and the payment boundary, addressing the concern that the original case study was overly qualitative and insufficiently clear about why samples were rejected. Specifically, for all rejected samples ($N = 21{,}557$), we assign each case to its lowest-scoring dimension, which we treat as the dominant failure dimension. We then report, for each dimension, its frequency among rejected samples, its category-wise distribution, and the mean scores for accepted and rejected samples, together with Cohen's d as an effect-size measure. The results show that BRF is the most frequent failure mode (53.8%) and also the dimension that best separates accepted from rejected samples ($d = 0.39$), suggesting that the main bottleneck lies in specification and constraint satisfaction. We also observe distinct category-level patterns: Portrait failures are more often associated with VEQ-Realism (23.7%) and VEQ-Clarity (17.7%), Product failures are dominated by BRF (64.7%), and Digital/Art failures are mainly driven by BRF (48.3%) and CNS (30.3%), highlighting the importance of cross-image consistency in such tasks. By contrast, VEQ-Aesthetic barely separates accepted and rejected outcomes ($d = -0.02$), further suggesting that requirement satisfaction and consistency, rather than aesthetic preference alone, more strongly define the payment boundary.

To further illustrate these failure modes, we provide representative failure examples in Table 13. These examples show that rejection may arise from missing required elements (BRF), insufficient clarity or realism, severe aesthetic degradation, text rendering errors, or poor consistency across mul-

Model	Total				Portrait		Product		Digital	
	Revenue ($k)	Share (%)	Task Acc. (%)	Deliv. Acc. (%)	Rev ($k)	Share (%)	Rev ($k)	Share (%)	Rev ($k)	Share (%)
Closed-Source Models										
Gemini-Banana	149.82	50.8	53.76	64.38	8.91	58.3	84.69	57.6	56.22	42.3
Gemini-Banana-Pro	199.24	67.5	73.85	74.96	0.03	0.2	143.99	98.0	55.21	41.6
GPT-Image-1	145.24	49.2	55.34	65.23	10.32	67.5	78.88	53.7	56.04	42.2
GPT-Image-1-Mini	147.93	50.1	56.25	65.30	9.62	63.0	83.10	56.5	55.20	41.6
Ideogram-v3	158.02	53.6	56.09	65.18	10.37	67.9	87.73	59.7	59.91	45.1
Imagen-4.0	183.27	62.1	70.19	80.45	13.71	89.7	62.84	42.8	106.72	80.4
Kling-v2	29.64	10.0	19.55	24.40	7.13	46.6	1.82	1.2	20.69	15.6
MJ-Relax-Edits	139.95	47.4	54.44	65.38	9.40	61.5	70.22	47.8	60.33	45.4
MJ-Relax-Imagine	136.33	46.2	55.58	66.11	10.01	65.5	69.61	47.4	56.71	42.7
SeedEdit-3.0-i2i	18.37	6.2	50.39	66.53	0.05	0.3	6.18	4.2	12.13	9.1
Seedream-3.0-t2i	144.60	49.0	55.07	66.19	10.23	66.9	73.99	50.3	60.38	45.5
Seedream-4.0	0.00	0.0	0.00	0.00	0.00	0.0	0.00	0.0	0.00	0.0
Open-Source Models										
FLUX-1.1-Pro	169.68	57.5	58.16	66.32	9.79	64.1	89.50	60.9	70.39	53.0
FLUX-Kontext-Pro	138.07	46.8	54.90	65.49	9.44	61.8	69.94	47.6	58.69	44.2
Qwen-Image	148.09	50.2	55.50	65.68	10.22	66.8	82.41	56.1	55.47	41.8
SD-3.5-Large	132.88	45.0	53.94	64.67	9.58	62.6	61.12	41.6	62.18	46.8

Table 9: Model performance on ServImage under the standard settlement scenario, **estimated using ServImageModel-predicted payment probabilities**. Revenue, Rev ($k), is the total contract value estimated from *predicted* accept/reject outcomes, and Share is its fraction of the overall contract value ($295k), while category shares are computed within each category. Task Acceptance and Deliverable Acceptance are the predicted proportions of tasks and deliverables approved under ServImageModel. (See Table 2 for results based on **human** payment-decision labels.) **Color:** 1st 2nd 3rd.

Judge	BRF vs Human	CNS vs Human	VEQ vs Human	Overall vs Human
GPT-5-mini	0.87	0.88	0.87	0.88
GPT-5	0.91	0.90	0.91	0.91
Gemini-3-Flash	0.80	0.89	0.84	0.85

Table 10: Agreement between automatic judges and human verification on 300 randomly sampled deliverables. For each dimension (BRF, VEQ, and CNS), we compute the agreement rate as the proportion of instances for which the judge's decision matches the human judgment. **Overall vs Human** is computed in the same way after aggregating the three dimensions into the final overall decision for each deliverable.

tiple generated images.

Pair of Judges	ρ_{BRF}	ρ_{VEQ}	ρ_{CNS}	ρ_{overall}
GPT-5-mini vs GPT-5	0.86	0.74	0.82	0.87
GPT-5-mini vs Gemini-3-Flash	0.77	0.69	0.77	0.79
GPT-5 vs Gemini-3-Flash	0.76	0.70	0.77	0.80

Table 11: Rank correlation between judge pairs over model rankings induced by each judge. For each dimension, ρ denotes the Spearman rank correlation coefficient computed from the full ranking of models under the corresponding judge. The overall ranking is obtained by recomputing ServImageScore with each judge and then measuring the Spearman correlation between the resulting overall rankings. In addition, we also examine Top-k overlap, defined as the proportion of shared models in the top-k positions between two judge-specific rankings.

Dimension	Overall (%)	Portrait (%)	Product (%)	Digital (%)	Acc. Score	Rej. Score	Cohen's d
BRF	53.8	40.8	64.7	48.3	3.48	2.82	0.39
VEQ: Clarity	5.1	17.7	2.4	3.0	4.69	4.54	0.26
VEQ: Realism	9.6	23.7	4.1	9.7	4.29	4.22	0.11
VEQ: Aesthetic	4.2	7.0	4.6	2.9	4.14	4.15	-0.02
VEQ: Text	5.2	5.5	4.6	5.8	4.33	4.17	0.18
CNS	22.1	5.4	19.6	30.3	3.66	3.42	0.22

Table 12: Systematic failure analysis over rejected samples ($N = 21{,}557$). For each rejected sample, the **dominant failure dimension** is defined as the lowest-scoring dimension for that sample. **Overall (%)** denotes the proportion of all rejected samples whose dominant failure falls into the corresponding dimension. **Portrait (%)**, **Product (%)**, and **Digital (%)** denote the proportion of rejected samples within each category whose dominant failure is the corresponding dimension. **Acc. Score** and **Rej. Score** are the mean scores of accepted and rejected samples, respectively, on that dimension. **Cohen's** d is computed as the standardized mean difference between accepted and rejected samples on the corresponding dimension, i.e., the difference in means divided by the pooled standard deviation.

Dimension	Model	Category	Task Description	Dominant Score	Failure Notes
BRF	kling-v2-images	Digital	Design 8 anime characters + expressions	2.0	Missing some of the 8 characters; missing four expressions; missing weekday-to-character mapping
VEQ: Clarity	qwen-text-to-image	Portrait	Photo with professional gray-toned outfit	4.0	BRF is perfect, but clarity, realism, and aesthetics are all 4 (tie $\rightarrow$ lowest dimension assignment)
VEQ: Realism	flux-1.1-pro	Digital	Bookworm animal reading scene	4.0	BRF (=4.5), Clarity (=5), Aesthetic (=5), but Realism (=4)
VEQ: Aesthetic	mj-relax-imagine_4	Digital	Historical characters from Tang/Song/Ming	0.0	Clarity (=4), Realism (=4), but Aesthetic (=0), indicating an extreme aesthetic failure
VEQ: Text	seedream-4.0	Digital	Animal cartoon brand mascot	2.0	Other dimensions are 4–5, but Text (=2), indicating a text rendering failure
CNS	flux-kontext-pro	Digital	Virtual fitness-coach IP design	2.5	BRF (=4.29), Clarity (=5), but CNS (=2.5), indicating poor consistency across 12 images

Table 13: Representative failure examples grouped by dominant failure dimension. **Dominant Score** denotes the score of the lowest-scoring dimension for the corresponding sample, which determines the failure label used in Table 12. When multiple dimensions tie for the minimum score, the example is assigned according to the predefined tie-breaking rule used in our analysis pipeline.