LAYERBENCH: BENCHMARKING LAYER DECOMPOSITION BEYOND RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Layer decomposition is fundamental to editable image manipulation, yet its evaluation remains misaligned with its purpose: a flat copy can achieve near-perfect reconstruction while being useless for editing. Existing protocols prioritize pixel fidelity, are tied to narrow distributions, and rarely validate model-based scores against human preferences. We introduce **LayerBench**, a comprehensive benchmark for layer decomposition that explicitly aligns evaluation with human preferences. LayerBench comprises 1.9K curated images spanning diverse domains, layer cardinalities, occlusion patterns, transparency effects, and semantic structures, together with a preference corpus of 13,979 expert pairwise comparisons used to train **LayerReward**. Its evaluation suite combines rule-based metrics for reconstruction, alignment, text preservation, and operational editability; model-based metrics from pretrained large vision-language models; and LayerReward, a five-dimensional evaluator that exceeds 80% overall pairwise accuracy and generalizes to out-of-distribution shadow separation. Beyond evaluation, LayerReward also serves as a reward signal for **LayerCompass**, a reward-guided post-training method included as part of LayerBench. Experiments show that reconstruction fidelity is necessary but insufficient for editability and human preference.

1 INTRODUCTION

Layer decomposition—recovering an ordered set of independently editable RGBA layers from a flattened RGB image—is a foundational capability for a wide range of image editing and graphic design applications (Yamaguchi, 2021; Suzuki et al., 2025). It underpins tasks such as object repositioning, text editing, background replacement, and effect manipulation, enabling users to modify individual elements of a design without re-creating it from scratch (Chen et al., 2025a). As generative models increasingly produce complex visual compositions, the ability to decompose these outputs into semantically meaningful and editable layers becomes ever more critical (Liu et al., 2025b; Wang et al., 2026). Despite rapid progress in the underlying models (Yin et al., 2026; Rowles et al., 2026; Wu et al., 2026), the field lacks a principled way to measure whether a decomposition is actually useful for its intended purpose.

The central difficulty is that layer decomposition is inherently underdetermined: a single RGB image can correspond to many distinct yet equally plausible layer stacks. For instance, a cast shadow may be assigned to the object that produces it or treated as an independent layer; a text block may be kept as one layer or split into multiple sub-layers. This ambiguity makes evaluation fundamentally different from standard reconstruction tasks. However, most existing protocols evaluate decomposition primarily through pixel-level reconstruction fidelity (Liu et al., 2025b; Lungu-Stan et al., 2026), which is problematic: a flat copy of the input achieves near-perfect reconstruction while being completely useless for editing. As we show, methods that lead in pixel reconstruction can rank lowest on LayerReward preference (LayerD anchored at 0.5), revealing a mismatch between reconstruction metrics and preferred decompositions.

Beyond reconstruction fidelity, current evaluation practices suffer from two limitations. First, reference-based protocols are tied to Crello (Yamaguchi, 2021), whose relatively narrow distribution and annotation conventions do not reflect the diversity of real-world decomposition scenarios, including photorealistic images, dense text layouts, and occlusion patterns. Second, while some recent works employ vision-language models (VLMs) as judges (Rowles et al., 2026; Lungu-Stan et al.,

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.

2026), these scores have not been systematically validated against human preferences or real editing outcomes. No established benchmark jointly assesses layer fidelity, compositional correctness, semantic organization, and practical editability while grounding its metrics in human judgment. As a result, progress across models with different conditioning requirements, output cardinalities, and target domains remains difficult to compare under a unified protocol.

To address this gap, we introduce **LayerBench**, a comprehensive benchmark for layer decomposition that explicitly aligns evaluation with human preferences. LayerBench comprises a curated data suite of 1.9K images spanning diverse visual domains and targeted challenges (photorealistic completion, dense text preservation, and shadow-aware decomposition); a unified evaluation suite that combines deterministic rule-based metrics—covering reconstruction correctness, layer alignment, text preservation, and operational editability—with model-based metrics from pretrained large vision-language models; and **LayerReward**, a five-dimensional evaluator trained on 13,979 expert pairwise comparisons, which exceeds 80% overall pairwise accuracy and generalizes to out-of-distribution shadow separation without shadow-specific training.

Beyond evaluation, LayerReward also serves as a reward signal for preference-aligned post-training. As part of LayerBench, we introduce **LayerCompass**, a reward-guided post-training method that optimizes a layer decomposition model against LayerReward with reinforcement learning. This closes the loop from evaluation to improvement: a human-aligned evaluator can be not only a diagnostic tool but also an actionable optimization objective. Experiments on LayerBench further show that reconstruction fidelity is necessary but insufficient, that human-aligned evaluation captures semantic errors that pixel-level scores miss, and that LayerReward disfavors the fragmentation induced by fine-grained over-segmentation even when isolated editability probes reward it.

The main contributions of this work are summarized as follows. We propose LayerBench, a comprehensive benchmark that harmonizes human-preference alignment with practical editability in layer decomposition through a diverse data suite, a unified rule- and model-based evaluation protocol, and LayerReward. LayerBench also includes LayerCompass as a reward-guided post-training component that uses LayerReward to improve decomposition quality. We believe the release of LayerBench, together with its preference annotations and trained reward model, will facilitate future research in editable layer decomposition and human-aligned generative modeling.

2 Related Works

Layer Decomposition. Layer decomposition aims to recover an ordered set of independently editable RGBA layers from a flattened RGB image. Accordion (Chen et al., 2025a) uses a vision-language model to plan a top-down decomposition and coordinates specialized vision modules to convert raster graphic designs into editable layers. LayerD (Suzuki et al., 2025) instead repeatedly extracts the topmost visible element, completes the residual canvas, and refines the recovered layers under graphic-specific priors. Qwen-Image-Layered (Yin et al., 2026) introduces an end-to-end diffusion model with an RGBA-aware VAE and a variable-layer transformer for decomposing one RGB image into a semantic RGBA stack. Subsequent systems broaden this setting through unified generation and decomposition (Lungu-Stan et al., 2026), large-scale masked-region modeling (Tang et al., 2026), multi-task transparency-aware post-training (Yu et al., 2025), VLM-scored reinforcement learning (Rowles et al., 2026), and box-guided two-stage decomposition pipelines such as SynLayers (Wu et al., 2026). RevealLayer (Wang et al., 2026) further targets occlusion-heavy natural images, while UniWorld-Design (Li et al., 2026) supports instruction-conditioned joint recovery of ordered, complete semantic layers. Despite this progress, methods differ substantially in conditioning requirements, output cardinality, target domain, and layer semantics, making their capabilities difficult to compare under a unified protocol.

Related work also considers color-based palette decomposition and mask- or target-conditioned extraction of a specified foreground and its completed background (Tan et al., 2016; Aksoy et al., 2017; Tan et al., 2018; Yang et al., 2025; Kang et al., 2025; Chen et al., 2026). In addition, multi-layer image generation frameworks synthesize layered visual content from scratch conditioned on text prompts or regional layouts (Zhang et al., 2023; Zhang & Agrawala, 2024; Pu et al., 2025; Huang et al., 2024), operating in a forward generative setting rather than image-to-layer decomposition. Since these methods do not recover a complete variable-length semantic RGBA layer stack from an unconstrained flattened image, they fall outside the primary scope of this work.

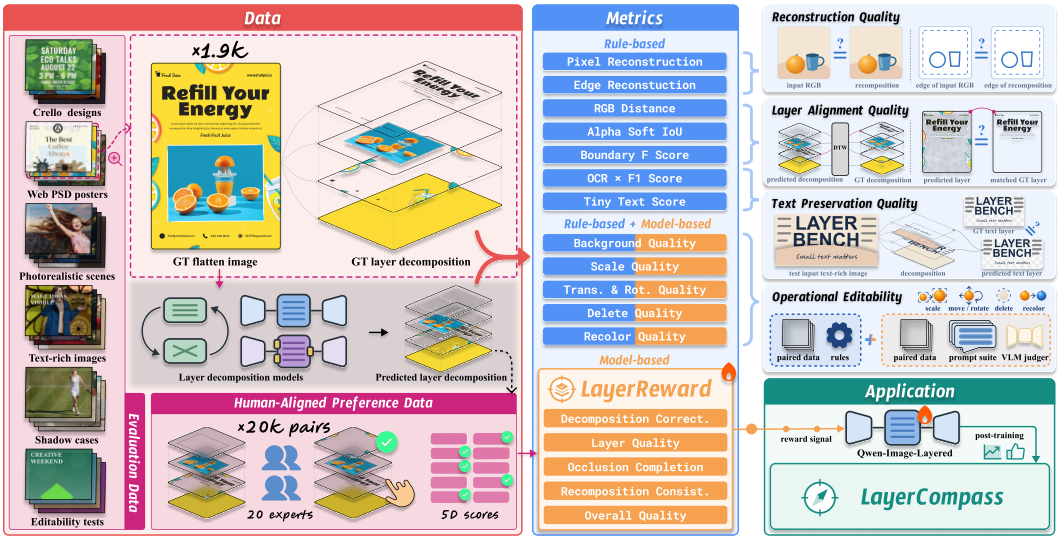


Figure 1: LayerBench evaluates predicted RGBA stacks with rule-based and human-aligned metrics. Expert preferences train LayerReward, which also guides LayerCompass post-training.

Evaluation of Layer Decomposition. Most existing protocols evaluate on Crello (Yamaguchi, 2021), whose source designs provide ordered ground-truth elements. LayerD uses order-aware Dynamic Time Warping with adjacent reference-layer merging and reports alpha-weighted RGB L_1 and Alpha SoftIoU; Qwen-Image-Layered adopts the same protocol (Suzuki et al., 2025; Yin et al., 2026). Stable-Layers instead uses best-match assignment and per-layer RGB L_1, while LaDe adds PSNR and a VLM judge (Rowles et al., 2026; Lungu-Stan et al., 2026). Broader protocols combine conventional fidelity measures with learned judgments: CLD evaluates layers, alpha masks, and recompositions and uses Q-Insight to assess semantic consistency, visual fidelity, and editability (Liu et al., 2025b); Stable-Layers defines a VLM rubric covering semantic separation, alpha quality, background completion, and content validity (Rowles et al., 2026). These protocols remain fragmented across datasets and scoring conventions; LayerBench targets this gap.

3 LayerBench

LayerBench unifies evaluation data, deterministic metrics, and human-preference-aligned assessment under a single protocol (Figure 1). It contains 1.9K curated images and LayerReward, trained on 13,979 preference pairs, which also guides LayerCompass.

3.1 Dataset Curation

As summarized in Figure 2, LayerBench contains 1,902 images from six complementary subsets. We select 500 clean, template-driven designs from the Crello test split and curate 500 online layered PSDs after quality filtering. To extend beyond poster-like graphics, we add 452 photorealistic images that require completion of hidden textures, illumination, and scene structure. The suite also includes 200 text-rich images (half with dense small-font text) and 100 editability-probe examples.

The remaining 150 examples form a shadow-focused subset for evaluating shadow ownership and effect-aware grouping. From paired shadow and shadow-free images (Liu et al., 2024), we recover an explicit black RGBA shadow layer and combine it with a manually verified semantic decomposition of the shadow-free image. We retain stacks whose recomposition matches the original shadow image; Appendix A.1 provides the full construction procedure. Together, the six subsets cover broad visual diversity and targeted challenges in amodal completion, text preservation, and editability.

3.2 Rule-Based Metrics and Editability Probes

Let $\mathcal{P} = (P_1, \ldots, P_N)$ and $\mathcal{G} = (G_1, \ldots, G_M)$ denote the predicted and ground-truth RGBA layer stacks, respectively, ordered from bottom to top. Our deterministic evaluation covers four dimensions: reconstruction correctness, layer alignment, text preservation, and operational editability. Unless otherwise specified, all scores lie in $[0, 1]$, with higher values indicating better performance. Full definitions and implementation details are provided in Appendix A.3.



Figure 2: Composition and layer-count distribution of the 1,902-image LayerBench data suite.

Reconstruction correctness. We alpha-composite the predicted stack and measure both pixel fidelity to the input and preservation of the internal edge structure in the reference stack. The latter prevents reconstruction quality from being dominated by a flattened copy of the input. We average the pixel and edge scores to obtain S_{recon}. If a model returns a single layer that nearly reproduces the input while the reference contains multiple layers, we treat it as a trivial flattened-image solution and set both reconstruction and alignment scores to zero.

Layer alignment. Because valid decompositions may differ in ordering and cardinality, we first group mutually non-overlapping layers on the same visual plane and then align the resulting sequences with Dynamic Time Warping. Each matched pair is scored by premultiplied RGB agreement $\overline{D}_{\text{rgb}}$, Alpha SoftIoU $\overline{S}_{\alpha}$, and boundary preservation $\overline{S}_{b}$. Given the optimal alignment path, the final score is

$$S_{\text{align}} = 0.4(1 - \overline{D}_{\text{rgb}}) + 0.4\overline{S}_{\alpha} + 0.2\overline{S}_{b}. \tag{1}$$

We additionally allow limited merging of adjacent predicted layers.

Text-layer correctness. Text requires readability and clean isolation. We use zero-to-many matching, excluding the bottommost background layer, and combine OCR character agreement with the Alpha F1 of the selected text support. On the Tiny Text subset, we instead score the designated small-text layer using appearance, Alpha SoftIoU, and boundary preservation. The reported Text score is the mean of the scores computed on the OCR and Tiny Text subsets.

Rule-based editability. We measure editability operationally by applying the same concrete editing operation to the predicted and reference layer stacks and comparing the resulting changes in their recompositions. This tests whether the predicted layers support the intended manipulation, rather than only matching a static reference. We treat the bottommost predicted and reference layers as their respective backgrounds. Each foreground reference target is matched to the non-background predicted layer with the highest Alpha SoftIoU,

$$\pi(t) = \arg\max_{j \geq 2} S_{\alpha}(P_j, G_t). \tag{2}$$

We keep this correspondence fixed across editability probes. Background quality measures visible and occluded-region completion, opacity, and boundary seams. For foreground objects and text, we apply identical scaling, translation–rotation, deletion, and recoloring operations to the matched predicted and reference layers. We compare their edit-induced change maps,

$$\Delta_P = I_P^{\text{edit}} - I_P^{\text{orig}}, \qquad \Delta_G = I_G^{\text{edit}} - I_G^{\text{orig}}. \tag{3}$$

Each operation combines change-color agreement, soft support overlap, and change-edge preservation. The overall editability score averages the background and four foreground editability probes:

$$S_{\text{edit,rule}} = \text{Mean}(S_{\text{bg}}, S_{\text{scale}}, S_{\text{trrot}}, S_{\text{delete}}, S_{\text{recolor}})\,. \tag{4}$$

3.3 VLM-Assisted Editability Judge

To capture subtle perceptual artifacts (e.g., blending seams, semantic bleeding) that analytical rules miss, we adopt a pretrained VLM judge on a 1–5 scale (Appendix A.5). Scores map to $[0, 1]$ via $(s - 1)/4$; S_{VLM} averages the five probes.

3.4 LAYERREWARD: PREFERENCE-ALIGNED EVALUATION

Rule-based metrics miss semantic organization, perception, and hidden-content plausibility; LayerReward adds five expert-aligned scores, detailed below.

3.4.1 PREFERENCE DATASET

We sample 8,000 source images from PrismLayersPro (Chen et al., 2025b) together with their diverse candidate decompositions generated under multiple model and inference configurations. To cover both substantial cross-model differences and subtler within-model variations, we vary available inference configurations, including random seeds and target layer counts. Candidates from the same source are randomly paired, yielding an initial pool of 20,000 distinct comparisons. These preference pairs are collected separately from the 1,902-image LayerBench evaluation suite.

Annotations are provided by 20 experts. For each pair, annotators compare decomposition correctness (content assignment and granularity), layer quality (appearance and alpha fidelity), occlusion completion (plausibility of hidden content), recomposition consistency (input reconstruction), and overall quality (accuracy and editability). For every dimension, they choose whether candidate A is better, candidate B is better, or the two are tied. The complete annotation rubric is provided in Appendix A.6. After iterative annotation and careful verification, we retain 13,979 high-quality preference pairs (12,979 train / 1,000 test; no image overlap).

3.4.2 REWARD MODEL TRAINING

We build LayerReward on Qwen3-VL-8B-Instruct (Bai et al., 2025a), leveraging its visual priors and reasoning ability to assess decomposition quality. Since its vision encoder accepts RGB inputs, we composite each RGBA layer onto a checkerboard background to obtain an RGB visualization. Given an image I and an M-layer decomposition $\mathcal{D} = \{L_i\}_{i=1}^{M}$, we extract the final-layer representation $\mathbf{h}(I, \mathcal{D})$ from Qwen3-VL and map it through a newly added linear regression head to a five-dimensional reward vector, with each entry corresponding to one of the five quality dimensions:

$$\mathbf{r}(I, \mathcal{D}) = W\mathbf{h}(I, \mathcal{D}) + \mathbf{b} \in \mathbb{R}^5, \tag{5}$$

Here, W and $\mathbf{b}$ are the trainable parameters of the regression head. We fine-tune the Qwen3-VL backbone with LoRA and jointly optimize the regression head using pairwise human preferences. For each training pair, we denote the two candidate decompositions as A and B and score them independently. Let r_A^k and r_B^k be their rewards for dimension k, with preference probability

$$p^k = \frac{\exp(r_A^k)}{\exp(r_A^k) + \exp(r_B^k)}. \tag{6}$$

The training objective is the binary cross-entropy averaged over the five reward dimensions:

$$\mathcal{L}_{\text{reward}} = -\frac{1}{5}\sum_{k=1}^{5}\left[y^k \log p^k + (1 - y^k)\log(1 - p^k)\right]. \tag{7}$$

We set y^k to 1 when A is preferred, 0 when B is preferred, and 0.5 for ties.

3.5 PREFERENCE-ALIGNED POST-TRAINING WITH LAYERREWARD

We test LayerReward as an optimization signal by post-training Qwen-Image-Layered with DiffusionNFT (Zheng et al., 2026). Following the randomized-cardinality setting of StableLayers (Rowles et al., 2026), we sample $N \sim \text{Uniform}\{3, 4, 5\}$.

For each source image and sampled N, the policy generates K candidates $\{\mathcal{P}_i\}_{i=1}^{K}$. LayerReward assigns each a score $s_i^{(d)}$ for each of the $D = 5$ dimensions. We normalize scores within the group

and average the dimension-wise advantages:

$$A_i^{(d)} = \frac{s_i^{(d)} - \mu^{(d)}}{\sigma^{(d)} + \epsilon}, \qquad A_i = \frac{1}{D}\sum_{d=1}^{D} A_i^{(d)}. \tag{8}$$

Following DiffusionNFT, we convert the aggregated advantage to an optimality probability

$$r_i = \frac{1}{2} + \frac{1}{2}\,\mathrm{clip}\left(\frac{A_i}{Z}, -1, 1\right) \in [0, 1], \tag{9}$$

where $Z > 0$ controls the reward scale. For a clean candidate $\mathbf{x}_0^i$, we sample a diffusion time t and noise $\boldsymbol{\epsilon}$, obtaining $\mathbf{x}_t^i = \alpha_t \mathbf{x}_0^i + \sigma_t \boldsymbol{\epsilon}$ and the velocity target $\mathbf{v} = \dot{\alpha}_t \mathbf{x}_0^i + \dot{\sigma}_t \boldsymbol{\epsilon}$. We then use the original DiffusionNFT implicit positive and negative velocity parameterizations,

$$\mathbf{v}_\theta^+ = (1-\beta)\mathbf{v}^{\mathrm{old}} + \beta \mathbf{v}_\theta, \qquad \mathbf{v}_\theta^- = (1+\beta)\mathbf{v}^{\mathrm{old}} - \beta \mathbf{v}_\theta, \tag{10}$$

and optimize its negative-aware flow-matching objective:

$$\mathcal{L}_{\mathrm{NFT}}(\theta) = \mathbb{E}_{i,t,\boldsymbol{\epsilon}}\left[r_i \left\| \mathbf{v}_\theta^+(\mathbf{x}_t^i, \mathbf{c}, t) - \mathbf{v} \right\|_2^2 + (1 - r_i) \left\| \mathbf{v}_\theta^-(\mathbf{x}_t^i, \mathbf{c}, t) - \mathbf{v} \right\|_2^2 \right], \tag{11}$$

where $\mathbf{c}$ contains the source image and the sampled layer count N.

4 Experiments

We evaluate four representative open-source layer decomposition methods: LayerD, Qwen-Image-Layered, Stable-Layers, and SynLayers, together with **LayerCompass**, obtained by post-training Qwen-Image-Layered for 200 steps using LayerReward (Section 3.5; Appendix A.8).

Evaluation protocol. For methods requiring an output layer count N, our primary *Fixed-K* setting uses the modal GT cardinality of each subset, avoiding privileged per-image information. We additionally report an oracle setting using exact per-image GT cardinality.

Since LayerReward is trained with a Bradley–Terry pairwise objective, we convert its outputs into preference probabilities relative to LayerD:

$$S_{\mathrm{LR}}^{(d)}(m) = \frac{1}{N}\sum_{i=1}^{N} \sigma\left(r_{m,i}^{(d)} - r_{\mathrm{LayerD},i}^{(d)} \right), \tag{12}$$

where LayerD is anchored at 0.5 and *LR Pref.* is the Overall score.

For the final score, we equally balance *benchmark-designed evaluation* and *human-preference-aligned evaluation*. The rule-based aggregate is a weighted mean of the four dimensions,

$$S_{\mathrm{Rule}} = \frac{1 \cdot S_{\mathrm{recon}} + 3 \cdot S_{\mathrm{align}} + 3 \cdot S_{\mathrm{text}} + 3 \cdot S_{\mathrm{edit,rule}}}{10}, \tag{13}$$

which downweights reconstruction relative to alignment, text, and editability. Combining S_{Rule}, S_{VLM}, and S_{LR} yields

$$S_{\mathrm{Final}} = \frac{1}{2}\left(\frac{S_{\mathrm{Rule}} + S_{\mathrm{VLM}}}{2} + S_{\mathrm{LR}} \right) = \frac{S_{\mathrm{Rule}} + S_{\mathrm{VLM}} + 2S_{\mathrm{LR}}}{4}. \tag{14}$$

4.1 Results and Analysis on LayerBench

Tables 1 and 2 reveal clear trade-offs among decomposition paradigms. LayerD performs best in pixel reconstruction (0.957; Table 1), while SynLayers achieves the strongest text (0.656) and editability aggregates (0.900 rule-based; 0.679 VLM; Table 2). Stable-Layers obtains stronger preference-aligned scores than Qwen-Image-Layered, but at the cost of reconstruction fidelity. These differences motivate combining benchmark-designed and preference-aligned evaluation.

LayerCompass consistently improves its Qwen-Image-Layered base model: the aggregate Rule, VLM, and LR Pref. scores increase from 0.671 to 0.734, 0.317 to 0.529, and 0.690 to 0.775,

Table 1: Per-metric results on LAYERBENCH under the unified protocol. Top: rule-based metrics; Bottom: VLM judge (1–5) and LayerReward preference probabilities relative to LayerD. Best results are shown in **bold**; second-best results are underlined (↓: lower is better).

	Decomposition					Text		Editability (rule)				
Model	Pix.	Edge	Prem.↓	αIoU	BoundF	OCR	Tiny	BG	SC	TR	DL	RC
LayerD	**0.957**	0.739	0.182	0.726	0.624	0.424	0.453	**0.809**	0.563	0.568	0.602	0.557
SynLayers	0.676	0.793	0.193	0.714	0.626	0.609	**0.703**	0.766	**0.930**	**0.923**	**0.928**	**0.951**
Stable-Layers	0.626	0.771	0.144	0.788	**0.702**	0.433	0.540	0.576	0.672	0.675	0.704	0.689
Qwen-Image-Layered	0.805	0.758	0.144	0.774	0.681	0.449	0.580	0.675	0.658	0.661	0.688	0.680
LayerCompass	0.832(+3%)	**0.801**(+6%)	**0.133**(-8%)	**0.794**(+3%)	0.697(+2%)	**0.620**(+38%)	0.592(+2%)	0.766(+14%)	0.756(+15%)	0.758(+15%)	0.783(+14%)	0.767(+13%)

	Editability VLM judge (1–5)					LayerReward preference probability				
Model	BG	SC	TR	DL	RC	Decomp.	LayerQ	Occl.	Recomp.	Overall
LayerD	1.09	3.32	1.72	2.84	1.88	0.500	0.500	0.500	0.500	0.500
SynLayers	3.27	**4.70**	**2.72**	**4.42**	**3.48**	0.610	0.632	0.614	0.525	0.629
Stable-Layers	2.56	3.42	1.66	3.82	1.80	0.734	0.719	0.695	0.577	0.738
Qwen-Image-Layered	1.96	2.69	1.55	3.52	1.61	0.684	0.680	0.660	0.558	0.690
LayerCompass	**3.80**(+94%)	3.69(+37%)	1.87(+21%)	4.24(+20%)	1.98(+23%)	**0.770**(+8.6 pp)	**0.753**(+7.3 pp)	**0.727**(+6.7 pp)	**0.596**(+3.8 pp)	**0.775**(+8.5 pp)

Table 2: Dimension-level aggregates and final LAYERBENCH scores. Best results are shown in **bold**; second-best results are underlined.

	Rule-Based Dimensions				Aggregate Scores			
Model	Recon.	Align.	Text	Edit.	Rule	VLM Edit.	LR Pref.	**Final**
LayerD	**0.848**	0.742	0.439	0.620	0.625	0.293	0.500	0.480
SynLayers	0.735	0.734	**0.656**	**0.900**	**0.760**	**0.679**	0.629	0.674
Stable-Layers	0.698	0.798	0.486	0.663	0.654	0.413	0.738	0.636
Qwen-Image-Layered	0.781	0.788	0.515	0.672	0.671	0.317	0.690	0.592
LayerCompass	0.817(+5%)	**0.804**(+2%)	0.606(+18%)	0.766(+14%)	0.734(+9%)	0.529(+67%)	**0.775**(+8.5 pp)	**0.703**(+19%)

respectively. Accordingly, the final score increases from 0.592 to 0.703, the highest among the evaluated methods, followed by SynLayers at 0.674. This lead holds for LayerReward weights $w_{LR} \in [0.4, 0.6]$ (Appendix A.11). Because S_{Final} assigns half of its weight to LayerReward, part of this lead reflects optimization against the same preference signal; the accompanying gains on reconstruction, text preservation, and both rule- and VLM-based editability provide independent evidence beyond the reward being optimized.

This improvement is further supported by the blind human A/B study: LayerCompass obtains a 76.4% preference share over Qwen-Image-Layered (95% CI [72.0%, 80.6%]) and wins 79.0% of pairs by strict majority (Appendix A.9), supporting that LayerReward-guided post-training improves perceptually meaningful decomposition quality.

Our Fixed-K protocol already provides N-conditioned models with the modal GT count of each subset. Relative to this baseline, oracle per-image GT counts change Final and LR Pref. by at most 0.012 in magnitude, while text scores slightly decrease; editability is unchanged because the photo and editability subsets reuse Fixed-K predictions whose layer counts already match GT (Table 3). Exact per-image cardinality therefore provides limited benefit beyond the dataset-level prior.

4.2 QUALITATIVE COMPARISON

As shown in Figure 4, qualitative comparisons illustrate representative strengths and failure modes. LayerD sometimes collapses to a trivial single-layer copy of the input and makes layer-assignment errors in complex layouts. Qwen-Image-Layered achieves meaningful semantic separation, but often adds spurious solid white background layers and bakes cast shadows into the background. Stable-Layers handles shadows more plausibly, yet likewise introduces redundant white background layers. SynLayers also mishandles shadow ownership, but produces finer granularity that better isolates text and graphic assets. In these cases, LayerCompass yields the most visually coherent stacks among the compared methods: cleaner shadow separation, fewer spurious backgrounds, and a better qualitative trade-off between granularity and correctness, even though SynLayers remains stronger on some quantitative text and editability metrics.

4.3 ALIGNMENT OF RULE-BASED METRICS WITH HUMAN PREFERENCES

To validate LayerBench's rule-based metrics, we sample 300 prediction pairs: 100 jointly evaluated for reconstruction and alignment, 100 for text, and 100 for editability. Three raters perform two-alternative forced-choice comparisons, with the strict majority as the human verdict. Table 4 reports

Table 3: Oracle layer-count evaluation. Fixed K uses subset modes, whereas GT uses per-sample counts. The photo and editability subsets reuse Fixed-K predictions, whose layer counts already match GT and are therefore left unchanged under the oracle setting.

Model	Setting	Recon.	Align.	Text	Edit.	Rule	VLM	LR Pref.	Final
Stable-Layers	Fixed K	0.698	0.798	0.486	0.663	0.654	0.413	0.738	0.636
	GT layers	0.703(+0.005)	0.802(+0.004)	0.476(−0.010)	0.663(0.000)	0.653(−0.001)	0.413(0.000)	0.745(+0.007)	0.639(+0.003)
Qwen-Image-Layered	Fixed K	0.781	0.788	0.515	0.672	0.671	0.317	0.690	0.592
	GT layers	0.790(+0.009)	0.790(+0.002)	0.491(−0.024)	0.672(0.000)	0.665(−0.006)	0.317(0.000)	0.702(+0.012)	0.596(+0.005)
LayerCompass	Fixed K	0.817	0.804	0.606	0.766	0.734	0.529	0.775	0.703
	GT layers	0.819(+0.002)	0.807(+0.003)	0.579(−0.027)	0.766(0.000)	0.727(−0.007)	0.529(0.000)	0.776(+0.001)	0.702(−0.001)

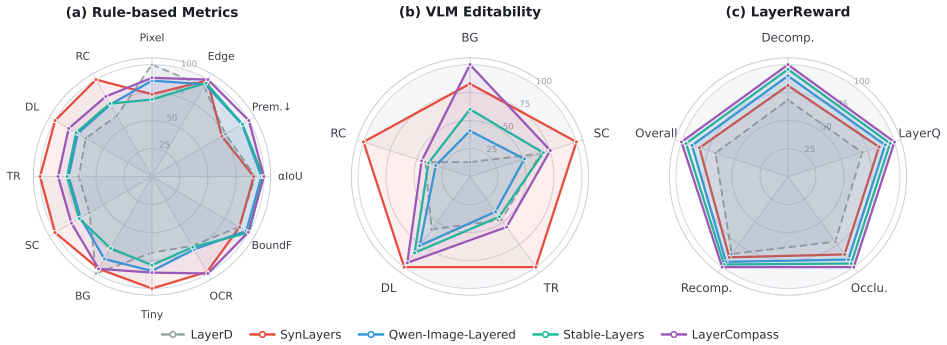


Figure 3: Multi-dimensional comparison of layer decomposition models on LayerBench. Scores are normalized per metric against the best-performing model; the premultiplied RGB distance is reversed so that larger values are better on all axes.

agreement between rule- and human-preferred predictions, bootstrap 95% confidence intervals, and pairwise inter-rater agreement. The 343 overall comparisons count dimension-level majority votes after excluding ties without a strict majority; reconstruction and alignment share the same 100-pair pool, so the total is not 300 independent samples.

Overall agreement is 75.2% (95% CI: 70.6–79.6%), above the 50% chance level. Editability (0.959) and alignment (0.811) track human judgments most closely, whereas reconstruction (0.631) and text (0.547) are less aligned because pixel and OCR proxies incompletely capture blending, soft shadows, and stylized text. Their lower inter-rater agreement (0.603 and 0.647) also indicates greater subjective ambiguity. Thus, rule-based metrics provide efficient, interpretable signals, while VLM-based editability and LayerReward capture complementary perceptual factors.

4.4 LayerReward Evaluation

We evaluate LayerReward on the held-out test set of 1,000 preference pairs (Table 5). It exceeds 80% overall accuracy (84.24%) and reaches over 80% on four of five dimensions (occlusion 90.61%; recomposition 74.61%; ablations in Appendix A.7). A blind five-model ranking study with four raters yields exactly the same model ordering as LayerReward (Appendix A.10).

We further test whether LayerReward recognizes an out-of-distribution preference for explicit shadow separation. On 500 held-out shadow-image pairs comparing a shadow-aware pipeline output with a direct decomposition of the same image, LayerReward—trained without shadow-domain data—favors the shadow-aware output on 71.4% of pairs for overall quality (mean preference probability 0.548), with the largest gains on decomposition correctness (73.4% win rate) and occlusion completion (69.6%); full results are in Appendix A.2.

5 Discussions and Insights

Complementary evaluation signals. LayerBench separates methods along the axes they target rather than collapsing them into a reconstruction-only ranking: LayerD leads pixel reconstruction, while other models trade reconstruction for stronger text isolation, editability, and/or preference-aligned quality. These axes are grounded in human judgments at three levels—rule-based metrics agree with majority votes on 75.2% of pairs, LayerReward reaches 84.24% overall accuracy on held-out preferences, and a blind study confirms LayerCompass gains over Qwen-Image-Layered. Reconstruction and editability are therefore not interchangeable: deterministic metrics capture pho-

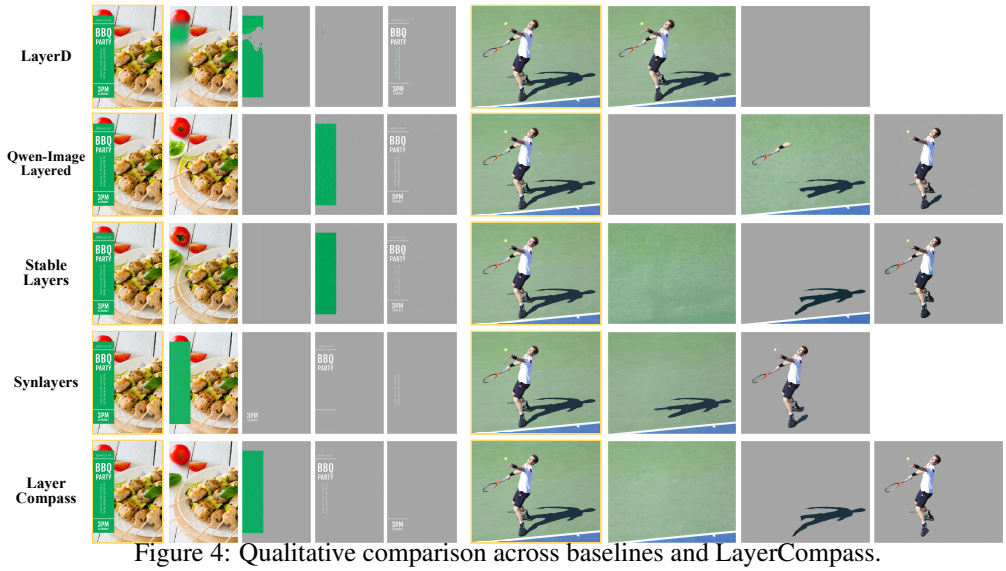


Figure 4: Qualitative comparison across baselines and LayerCompass.

Table 4: Alignment of rule-based metrics with human major-ity votes. Agreement is mean pairwise inter-rater agreement.

Dimension	#Pairs	Acc.	95% CI	Agreement
Reconstruction	65	0.631	[0.508, 0.754]	0.603
Layer Alignment	95	0.811	[0.726, 0.884]	0.810
Text	86	0.547	[0.442, 0.651]	0.647
Editability	97	0.959	[0.918, 0.990]	0.827
Overall	343	0.752	[0.706, 0.796]	0.722

Table 5: LAYERREWARD pairwise accuracy on the held-out test set.

Reward Dimension	Accuracy
Decomposition Correctness	83.69%
Layer Quality	88.18%
Occlusion Completion	90.61%
Recomposition Consistency	74.61%
Overall Quality	84.24%

tometric and structural properties, while LayerReward emphasizes semantic errors, broken occlusions, and missing content.

Rewards shape decomposition behavior. Stable-Layers and LayerCompass start from the same Qwen-Image-Layered base, so their comparison largely isolates the effect of the reward signal, although the RL algorithms also differ (GRPO (Liu et al., 2025a) vs. DiffusionNFT). A VLM-rubric reward raises judged editability but lowers reconstruction, whereas the multi-dimensional LayerReward improves LayerCompass on model-free structural, text, and editing metrics as well. Post-training should therefore balance complementary criteria rather than a single perceptual score.

Semantic granularity. Fine-grained slicing improves isolated editability probes, but LayerReward, trained on expert preferences, disfavors the fragmentation observed in SynLayers. A better target is semantic co-planarity: resolve depth and transparency across visual planes while grouping co-planar content into shared RGBA layers. When finer local control is needed, it can be recovered with lightweight operations (e.g., connected components).

Open directions. Important challenges remain: inferring layer count rather than conditioning on N; reducing autoencoder reconstruction error and full-stack sampling cost; improving hidden-region completion without over-segmenting co-planar content; and making layer-wise reinforcement learning cheaper and less dependent on fixed cardinality. The modest benefit of oracle layer counts (Table 3) further suggests that cardinality alone does not close the preference gap.

6 CONCLUSION

We presented LayerBench, a unified benchmark for layer decomposition that evaluates reconstruction, layer alignment, text preservation, operational editability, and human-preference-aligned quality under a single protocol. Central to this suite is LayerReward, trained on expert pairwise comparisons: it exceeds 80% overall pairwise accuracy, generalizes to out-of-distribution shadow separation (Appendix A.2), and supplies the reward for LayerCompass, which improves Qwen-Image-Layered on every reported LayerBench dimension. Experiments across sequential and generative methods show that reconstruction-only rankings do not coincide with editability or expert preference. We hope LayerBench, together with LayerReward and LayerCompass, will provide a lasting shared standard that makes progress in editable layer decomposition more comparable and more tightly aligned with human judgment.

AI USE STATEMENT

Generative AI tools, including large language models (LLMs), were used to assist with manuscript writing and language refinement, including improvements to clarity, grammar, wording, and presentation. Pretrained vision-language models (VLMs) were also used as explicit methodological components of this work, including model-based evaluation and reward modeling, as described in the corresponding sections of the paper. All AI-assisted writing and model-based outputs used in the study were reviewed and validated by the authors. The authors take full responsibility for the accuracy, originality, experimental results, and final content of this work.

ETHICS STATEMENT

Data sources and licensing. Except for the online PSD subset, all images in LayerBench are drawn from publicly available open-source datasets, or are derived from them through our post-processing. For the online PSD subset, we will release at least the source URLs together with our processing code. All data are used for non-commercial research purposes only.

Human annotation. All preference labels and human studies were provided by expert annotators, who were informed of the purpose of the study. The annotation task involves only judging the quality of layer decompositions and collects no personal information about the annotators. We will release the annotated preference dataset upon publication.

REPRODUCIBILITY STATEMENT

As a benchmark, LayerBench places a strong emphasis on reproducibility. Throughout inference and evaluation, we fix random seeds wherever possible. Metric definitions and evaluation protocols are described in Section 3 and Appendix A.3, and training details of LayerReward and LayerCompass are provided in Appendix A.8. Upon publication, we will fully open-source the benchmark, evaluation code, LayerReward and LayerCompass models, and their weights to ensure that all results can be reproduced.

REFERENCES

Yağiz Aksoy, Tunç Ozan Aydin, Aljoša Smolić, and Marc Pollefeys. Unmixing-based soft color segmentation for image manipulation. *ACM TOG*, 2017.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

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

Ryan D Burgert, Brian L Price, Jason Kuen, Yijun Li, and Michael S Ryoo. Magick: A large-scale captioned dataset from matting generated images using chroma keying. In *CVPR*, 2024.

Jingxi Chen, Yixiao Zhang, Xiaoye Qian, Zongxia Li, Cornelia Fermuller, Caren Chen, and Yiannis Aloimonos. From inpainting to layer decomposition: Repurposing generative inpainting models for image layer decomposition. In *CVPR*, 2026.

Jingye Chen, Zhaowen Wang, Nanxuan Zhao, Li Zhang, Difan Liu, Jimei Yang, and Qifeng Chen. Rethinking layered graphic design generation with a top-down approach. In *ICCV*, 2025a.

Junwen Chen, Heyang Jiang, Yanbin Wang, Keming Wu, Ji Li, Chao Zhang, Keiji Yanai, Dong Chen, and Yuhui Yuan. Prismlayers: Open data for high-quality multi-layer transparent image generative models. *arXiv preprint arXiv:2505.22523*, 2025b.

Runhui Huang, Kaixin Cai, Jianhua Han, Xiaodan Liang, Renjing Pei, Guansong Lu, Songcen Xu, Wei Zhang, and Hang Xu. Layerdiff: Exploring text-guided multi-layered composable image synthesis via layer-collaborative diffusion model. In *ECCV*, 2024.

Kyoungkook Kang, Gyujin Sim, Geonung Kim, Donguk Kim, Seungho Nam, and Sunghyun Cho. Layeringdiff: Layered image synthesis via generation, then disassembly with generative knowledge. *arXiv preprint arXiv:2501.01197*, 2025.

Zongjian Li, Zhiyuan Yan, Chenxu Bai, Chen Chen, Haoxiang Sun, Shaodong Wang, Feize Wu, Shenghai Yuan, Bin Lin, Zheyuan Liu, et al. Uniworld-design: From pixel generation to layer-native design. *arXiv preprint arXiv:2608.03971*, 2026.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. In *NeurIPS*, 2025a.

Qingyang Liu, Junqi You, Jianting Wang, Xinhao Tao, Bo Zhang, and Li Niu. Shadow generation for composite image using diffusion model. In *CVPR*, 2024.

Zihao Liu, Zunnan Xu, Shi Shu, Jun Zhou, Ruicheng Zhang, Zhenchao Tang, and Xiu Li. Controllable layer decomposition for reversible multi-layer image generation. *arXiv preprint arXiv:2511.16249*, 2025b.

Vlad-Constantin Lungu-Stan, Ionuț Mironică, and Mariana-Iuliana Georgescu. Lade: Unified multi-layered graphic media generation and decomposition. In *CVPR*, 2026.

Yifan Pu, Yiming Zhao, Zhicong Tang, Ruihong Yin, Haoxing Ye, Yuhui Yuan, Dong Chen, Jianmin Bao, Sirui Zhang, Yanbin Wang, et al. Art: Anonymous region transformer for variable multi-layer transparent image generation. In *CVPR*, 2025.

Ciara Rowles, Reshinth Adithyan, Nikhil Pinnaparaju, Vikram Voleti, and Mark Boss. Stable-layers: Fine-tuning image layer decomposition models with vlm-scored reinforcement learning. *arXiv preprint arXiv:2605.30257*, 2026.

Tomoyuki Suzuki, Kang-Jun Liu, Naoto Inoue, and Kota Yamaguchi. LayerD: Decomposing raster graphic designs into layers. In *ICCV*, 2025.

Jianchao Tan, Jyh-Ming Lien, and Yotam Gingold. Decomposing images into layers via RGB-space geometry. *ACM TOG*, 2016.

Jianchao Tan, Jose Echevarria, and Yotam Gingold. Efficient palette-based decomposition and recoloring of images via RGBXY-space geometry. *ACM TOG*, 2018.

Zhicong Tang, Jingye Chen, Zhao Zhang, Mohan Zhou, Yuchi Liu, Yifan Pu, Yalong Bai, Ethan Smith, and Yuhui Yuan. Masked region transformer for layered image generation and editing at scale. In *CVPR*, 2026.

Binhao Wang, Shihao Zhao, Bo Cheng, Qiuyu Ji, Yuhang Ma, Liebucha Wu, Shanyuan Liu, Dawei Leng, and Yuhui Yin. Reveallayer: Disentangling hidden and visible layers via occlusion-aware image decomposition. *arXiv preprint arXiv:2605.11818*, 2026.

Kam Man Wu, Haolin Yang, Qingyu Chen, Yihu Tang, Jingye Chen, and Qifeng Chen. Does synthetic layered design data benefit layered design decomposition? *arXiv preprint arXiv:2605.15167*, 2026.

Kota Yamaguchi. Canvasvae: Learning to generate vector graphic documents. In *ICCV*, 2021.

Jinrui Yang, Qing Liu, Yijun Li, Soo Ye Kim, Daniil Pakhomov, Mengwei Ren, Jianming Zhang, Zhe Lin, Cihang Xie, and Yuyin Zhou. Generative image layer decomposition with visual effects. In *CVPR*, 2025.

Shengming Yin, Zekai Zhang, Zecheng Tang, Kaiyuan Gao, Xiao Xu, Kun Yan, Jiahao Li, Yilei Chen, Yuxiang Chen, Heung-Yeung Shum, et al. Qwen-image-layered: Towards inherent editability via layer decomposition. In *CVPR*, 2026.

Hao Yu, Jinglin Wang, Jiabo Zhan, Rui Chen, Zile Wang, Huaisong Zhang, Hongyu Li, Xinrui Chen, Yongxian Wei, and Chun Yuan. Omnialpha: Aligning transparency-aware generation via multi-task unified reinforcement learning. *arXiv preprint arXiv:2511.20211*, 2025.

Lvmin Zhang and Maneesh Agrawala. Transparent image layer diffusion using latent transparency. *ACM TOG*, 2024.

Xinyang Zhang, Wentian Zhao, Xin Lu, and Jeff Chien. Text2layer: Layered image generation using latent diffusion model. *arXiv preprint arXiv:2307.09781*, 2023.

Kaiwen Zheng, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon, et al. Diffusionnft: Online diffusion reinforcement with forward process. In *ICLR*, 2026.

A Appendix

A.1 Shadow-Focused Subset Construction

Existing methods often handle cast shadows inconsistently: although a foreground object may be successfully extracted, its associated shadow remains baked into the background layer. Such results break the semantic association between an object and its visual effects, leading to implausible decompositions and limiting subsequent object repositioning or removal. To explicitly evaluate this behavior, we construct paired reference decompositions that preserve the intended association between foreground objects and their cast shadows.

We build this subset from an existing dataset of paired shadow images I^{sh} and corresponding shadow-free images I^{sf}. We represent each cast shadow as a canonical black RGBA layer $L^{\mathrm{sh}} = (\mathbf{0}, \alpha^{\mathrm{sh}})$, whose alpha channel encodes its spatial support and continuously varying opacity. Under standard alpha compositing, the observed shadow image satisfies

$$I^{\mathrm{sh}}(x) = \alpha^{\mathrm{sh}}(x)\mathbf{0} + \big(1 - \alpha^{\mathrm{sh}}(x)\big) I^{\mathrm{sf}}(x). \tag{15}$$

We recover the shadow opacity from the luminance ratio at each pixel,

$$\alpha^{\mathrm{sh}}(x) = \mathrm{clip}\left(1 - \frac{\mathcal{Y}(I^{\mathrm{sh}}(x))}{\max(\mathcal{Y}(I^{\mathrm{sf}}(x)), \epsilon)}, 0, 1\right), \tag{16}$$

where $\mathcal{Y}(\cdot)$ converts an RGB value to luminance and ϵ prevents division by values close to zero in dark regions.

We then apply Qwen-Image-Layered to the shadow-free images and manually retain only decompositions with coherent object separation and plausible completed content. The recovered shadow layer and selected semantic decomposition are combined to form the complete reference decomposition of the original shadow image. Finally, we verify that recompositing the combined stack reproduces the paired shadow image, yielding an explicit and independently editable shadow layer. Figure 5 summarizes the construction pipeline.

A.2 Shadow-Aware Preference Study

We further use this subset to test whether LayerReward generalizes to a previously unseen decomposition preference. For each of 500 source shadow images, we construct a paired comparison between two outputs. The *shadow-aware* output is produced by the pipeline above: we decompose the corresponding shadow-free image, recover an explicit RGBA shadow layer, and combine the two into a decomposition of the original shadow image. The *direct* output is obtained by applying the same decomposition model directly to the original shadow image, without explicit shadow extraction or reassembly; consequently, the cast shadow may remain entangled with the background or foreground content.

Human annotators consistently prefer the shadow-aware outputs, indicating that explicit shadow separation better matches human expectations of semantic and editable layers. Importantly, LayerReward was trained on neither shadow-domain data nor these comparison pairs. Nevertheless, its

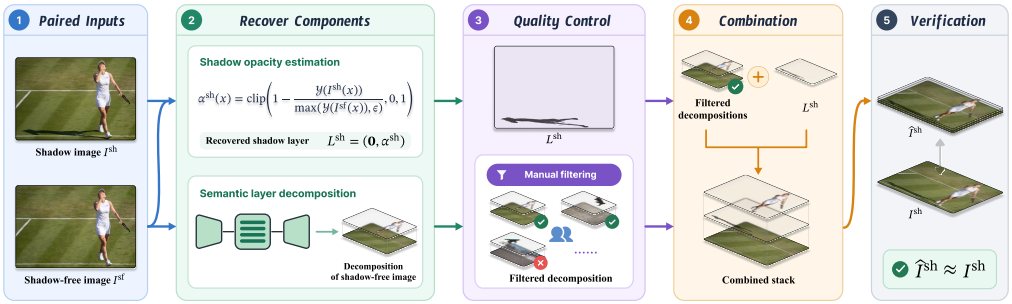


Figure 5: Construction of the shadow-focused reference decompositions. Starting from paired shadow and shadow-free images, we recover an explicit black RGBA shadow layer from their luminance ratio and decompose the shadow-free image into semantic layers. After manual quality filtering, the semantic stack and shadow layer are combined, and the resulting reference decomposition is verified by recomposition against the original shadow image.

Table 6: LayerReward preferences between shadow-aware pipeline outputs and direct decompositions on 500 image pairs. The reward model was not trained on shadow-domain data.

	Mean Score				
Reward Dimension	Shadow-Aware	Direct	Δ	Win Rate	$P(\text{prefer})$
Decomposition Correctness	-0.51	-0.68	$+0.17$	73.4%	0.541
Layer Quality	-0.67	-0.73	$+0.06$	59.8%	0.516
Occlusion Completion	$+2.19$	$+2.02$	$+0.18$	69.6%	0.543
Recomposition Consistency	$+4.58$	$+4.51$	$+0.07$	67.2%	0.518
Overall Quality	-1.75	-1.95	$+0.20$	71.4%	0.548

judgments agree with the human preference across all five reward dimensions (Table 6). For overall quality, it favors the shadow-aware output on 71.4% of pairs, with a mean preference probability of 0.548. This agreement provides two mutually supporting findings: the reward model generalizes to an out-of-distribution structural preference, and the shadow-aware construction yields decompositions that are genuinely preferred by humans rather than merely optimized for a task-specific metric.

The largest improvements occur in *Decomposition Correctness* ($\Delta = +0.17$, 73.4% win rate) and *Occlusion Completion* ($\Delta = +0.18$, 69.6%), the dimensions most directly related to structural organization and hidden-content completion. Gains in *Layer Quality* and *Recomposition Consistency* are smaller, as these properties are largely preserved by both variants. This pattern suggests that LayerReward recognizes the semantic benefit of assigning cast shadows to explicit layers, rather than responding only to superficial appearance differences.

A.3 Detailed Rule-Based Metrics

Let $\mathcal{P} = (P_1, \ldots, P_N)$ and $\mathcal{G} = (G_1, \ldots, G_M)$ denote the predicted and reference RGBA layer stacks, ordered from bottom to top.

Reconstruction correctness. We alpha-composite all predicted layers over a white canvas to obtain $\widehat{I} = \text{Comp}(\mathcal{P})$. Pixel-level fidelity is measured by

$$D_{\text{pixel}} = \frac{1}{3HW} \sum_{x,y,c} \left| \widehat{I}_{xyc} - I_{xyc} \right|, \qquad S_{\text{pixel}} = \text{clip}\left(1 - \frac{D_{\text{pixel}}}{0.08}, 0, 1\right). \tag{17}$$

To measure whether the predicted stack preserves the internal edge structure of the reference decomposition, for a layer $L_i = (C_i, \alpha_i)$ we define

$$E(L_i) = \text{clip}\big(\lambda_\alpha \mathcal{N}(\|\nabla \alpha_i\|_2) + \lambda_{\text{rgb}} \alpha_i \mathcal{N}(\|\nabla \bar{C}_i\|_2), 0, 1\big), \tag{18}$$

where $\bar{C}_i$ is the grayscale appearance and $\mathcal{N}$ denotes layer-wise min–max normalization. We aggregate layer edges using $E(\mathcal{L}) = \max_{L_i \in \mathcal{L}} E(L_i)$ and threshold the result to obtain $B(\mathcal{L})$. With

dilation operator $\mathcal{D}_r$, edge precision, recall, and F-score are

$$P_e = \frac{|B_P \cap \mathcal{D}_r(B_G)|}{|B_P| + \epsilon}, \quad R_e = \frac{|B_G \cap \mathcal{D}_r(B_P)|}{|B_G| + \epsilon}, \quad S_{\text{edge}} = \frac{2P_e R_e}{P_e + R_e + \epsilon}. \tag{19}$$

We use $\lambda_\alpha = 1$, $\lambda_{\text{rgb}} = 0.2$, edge threshold 0.1, and dilation radius $r = 2$, and define $S_{\text{recon}} = (S_{\text{pixel}} + S_{\text{edge}})/2$.

Layer alignment. We group mutually non-overlapping layers on the same visual plane through a top-down scan and align the resulting sequences using Dynamic Time Warping (DTW). For a predicted/reference pair (P, G), the matching cost is

$$d(P, G) = 0.4 D_{\text{rgb}}(P, G) + 0.4\big(1 - S_\alpha(P, G)\big) + 0.2\big(1 - S_b(P, G)\big), \tag{20}$$

where

$$D_{\text{rgb}}(P, G) = \frac{\sum_{x,y,c} |\alpha_P C_{P,c} - \alpha_G C_{G,c}|}{3 \sum_{x,y} \max(\alpha_P, \alpha_G) + \epsilon}, \quad S_\alpha(P, G) = \frac{\sum_{x,y} \min(\alpha_P, \alpha_G)}{\sum_{x,y} \max(\alpha_P, \alpha_G) + \epsilon}. \tag{21}$$

S_b averages the F-scores of alpha-mask boundaries and visible layer edges, both computed with dilation tolerance. Given the optimal DTW path $\mathcal{Q}$, we average the three terms along the path and obtain

$$S_{\text{align}} = 0.4(1 - \overline{D}_{\text{rgb}}) + 0.4\overline{S}_\alpha + 0.2\overline{S}_b. \tag{22}$$

To tolerate mild over-segmentation, we greedily allow up to three adjacent predicted-layer merges and retain the best score. If a model outputs only one layer that nearly reproduces the input ($D_{\text{pixel}} \leq 0.01$) while the reference contains multiple layers, we set both S_{recon} and S_{align} to zero.

Text-layer correctness. Because a text region may be distributed across multiple predictions, we use zero-to-many matching based on soft alpha overlap and edge agreement, excluding the bottommost predicted layer. The selected predicted and reference text layers are rendered for OCR using adaptive contrasting backgrounds. After Unicode normalization, case folding, punctuation standardization, whitespace removal, and reading-order sorting, character similarity is

$$S_{\text{char}} = 1 - \frac{d_{\text{Lev}}(T_P, T_G)}{\max(|T_P|, |T_G|, 1)}. \tag{23}$$

Let α_P^{text} and α_G^{text} denote the aggregated alpha canvases of the selected text layers. We compute soft alpha precision, recall, and F1 as

$$O_\alpha = \sum_{x,y} \min(\alpha_P^{\text{text}}, \alpha_G^{\text{text}}), \quad P_\alpha = \frac{O_\alpha}{\sum_{x,y} \alpha_P^{\text{text}} + \epsilon}, \quad R_\alpha = \frac{O_\alpha}{\sum_{x,y} \alpha_G^{\text{text}} + \epsilon}, \tag{24}$$

$$F_\alpha = \frac{2P_\alpha R_\alpha}{P_\alpha + R_\alpha + \epsilon}, \qquad S_{\text{ocr}} = S_{\text{char}} F_\alpha. \tag{25}$$

This term prevents readable text embedded in an impure or background-like layer from receiving a high score. For the designated small-text target G_{tiny}, we select the non-background prediction with the highest Alpha SoftIoU and combine support-normalized premultiplied color agreement, Alpha SoftIoU, and boundary preservation:

$$S_{\text{tiny}} = 0.4S_C + 0.4S_\alpha + 0.2S_b, \qquad S_{\text{text}} = \frac{S_{\text{ocr}} + S_{\text{tiny}}}{2}. \tag{26}$$

Rule-based editability. We designate the bottommost predicted layer P_1 and reference layer G_1 as their respective backgrounds. For each non-background reference target G_t, we independently select the non-background predicted layer with maximum Alpha SoftIoU,

$$\pi(t) = \arg \max_{j \in \{2,\ldots,N\}} S_\alpha(P_j, G_t), \tag{27}$$

and keep this correspondence fixed for all editing operations.

Background quality is evaluated using the bottommost predicted and reference layers, separately measuring visible and foreground-occluded regions together with opacity and seam quality:

$$S_{\text{bg}} = 0.15S_{\text{opacity}} + 0.25S_{\text{visible}} + 0.45S_{\text{hidden}} + 0.15S_{\text{seam}}. \tag{28}$$

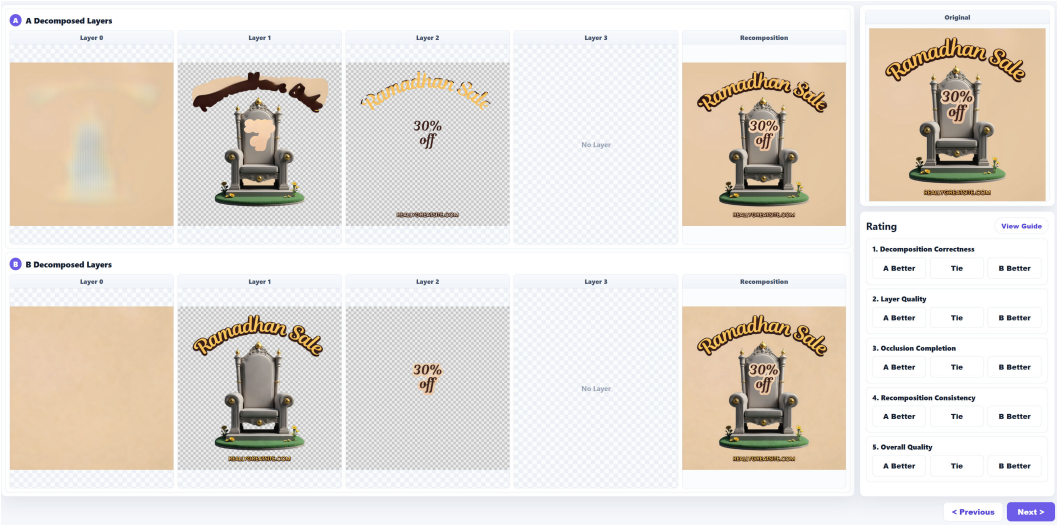


Figure 6: LayerReward annotation interface. Annotators compare two candidate decompositions and assign pairwise preferences along the five evaluation dimensions.

The higher weight on S_{hidden} emphasizes amodal completion, while S_{seam} penalizes foreground ghosts and discontinuities near the occlusion boundary.

For foreground object and text targets, we evaluate scaling, translation, rotation, deletion, and recoloring. The same transformation and reference-based pivot are applied to the matched predicted and reference layers. We compare the edit-induced change maps

$$\Delta_P = I_P^{\text{edit}} - I_P^{\text{orig}}, \qquad \Delta_G = I_G^{\text{edit}} - I_G^{\text{orig}}. \tag{29}$$

Within the union $\mathcal{R}$ of the target supports before and after editing, we measure change-color agreement, soft support overlap, and change-edge F-score:

$$S_\Delta = \text{clip}\left(1 - \frac{\sum_{(x,y)\in\mathcal{R},c} |\Delta_{P,xyc} - \Delta_{G,xyc}|}{3|\mathcal{R}| + \epsilon}, 0, 1\right), \tag{30}$$

$$M_P(x,y) = \max_c |\Delta_{P,xyc}|, \quad M_G(x,y) = \max_c |\Delta_{G,xyc}|, \quad S_{\text{sup}} = \frac{\sum_{\mathcal{R}} \min(M_P, M_G)}{\sum_{\mathcal{R}} \max(M_P, M_G) + \epsilon}, \tag{31}$$

$$S_{\text{op,rule}} = \frac{S_\Delta + S_{\text{sup}} + S_{\Delta e}}{3}. \tag{32}$$

If no non-background predicted layer is available, the corresponding target is left unmatched and receives a zero operation score. Finally,

$$S_{\text{edit,rule}} = \text{Mean}(S_{\text{bg}}, S_{\text{scale}}, S_{\text{trrot}}, S_{\text{delete}}, S_{\text{recolor}}). \tag{33}$$

A.4 Ground-truth construction

The 452 photorealistic examples are constructed by compositing transparent foreground mattes from MAGICK (Burgert et al., 2024) with curated natural-image backgrounds. The source background and foreground components are retained as the ground-truth RGBA layers. The 100 editability examples use a fixed three-component schema—background, object, and text—to provide unambiguous targets for controlled editing operations. No evaluated decomposition model is used to construct these ground-truth layers.

A.5 VLM Judge Prompts for Editability Evaluation

The editability judge is Qwen3-VL-8B-Instruct. All probe images are exported at the ground-truth resolution, and predicted layers are matched to GT layers with the same overlap-order matching



Figure 7: Three representative LayerReward annotation examples illustrating different pairwise preferences across the five evaluation dimensions.

used by the rule-based editability metric. For each sample the judge performs at most 13 calls: one background probe, one captioning call per target layer (the primary object layer and the text layer), and one judgment per probe operation per target layer (five operations $\times$ two target layers). Calls involving a target layer that cannot be matched in the prediction are skipped. The evaluation is two-stage: the model first captions each isolated target layer (composited over a neutral gray board), and the caption is injected into the subsequent judge prompt to unambiguously ground the edit target.

Background probe. The judge is shown the original poster, the GT background with the foreground removed, and the predicted background:

```
Image 1 is the original poster, Image 2 is the correct GT background with
foreground removed, and Image 3 is the predicted background. Judge Image 3
against Image 2, especially missing content, foreground ghosts, seams, blur,
and invented structure. Score 5: essentially correct; 4: minor defect;
```

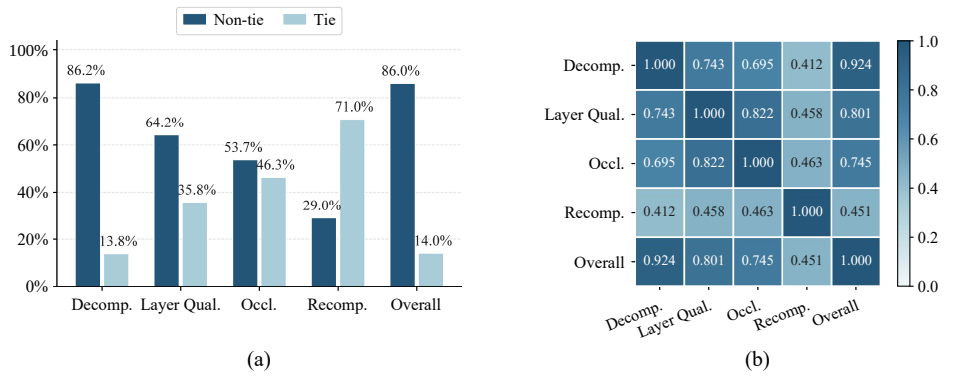


Figure 8: Analysis of pairwise preference annotations. (a) Non-tie and tie ratios across evaluation dimensions. (b) Pearson correlation heatmap of evaluation dimensions.

```
3: noticeable defect; 2: major defect; 1: unusable or mostly wrong.
Return JSON only: {"score": 1, "reason": "brief reason"}.
```

Stage 1: target captioning.

```
Describe only the visible content of this isolated layer in at most 12 words.
If it is text, transcribe the text. Return JSON only: {"caption": "..."}.
```

Stage 2: operation judging. The same edit is applied to the GT and the predicted target layer with an identical pivot (the GT bounding-box center), and the judge is shown the original poster, the GT-edited composite, and the prediction-edited composite. Each operation is verbalized as follows: *scale*: "scale the target to 75% about its center"; *translate*: "move the target right and upward"; *rotate*: "rotate the target 15 degrees"; *delete*: "delete the target"; *recolor*: "recolor the target while preserving its shape and shading". The judge prompt template is (line breaks added for readability):

```
Target: "<caption from stage 1>". Intended edit: <operation>. Image 1 is the
original poster, Image 2 is the correct GT-edited result, and Image 3 is the
candidate result. Judge only whether Image 3 performs the intended edit as
cleanly as Image 2, including residue, halos, broken edges, duplicated
content, or unintended changes. Score 5: correct and clean; 4: correct with a
minor artifact; 3: partly correct or clearly imperfect; 2: mostly wrong with
severe artifacts; 1: failed edit or wrong content changed.
Return JSON only: {"score": 1, "reason": "brief reason"}.
```

Output parsing and aggregation. The judge must return a JSON object with an integer score and a brief reason; responses that fail to parse are re-queried once with the suffix "Return one valid JSON object only.", and samples that still fail are discarded. Scores are averaged over the two target layers per operation, and *translate* and *rotate* are averaged into a single translation–rotation score. The reported VLM editability score is the mean of the five operation scores (background, scale, translation–rotation, delete, recolor), each normalized from the 1–5 scale to $[0, 1]$ via $(s - 1)/4$.

A.6 Preference Annotation Protocol

For each candidate pair, annotators compare the two decompositions according to the following five dimensions:

1. **Decomposition Correctness**: whether image content is assigned to appropriate layers at a reasonable granularity, without missing or duplicated elements, incorrect assignments, or cross-layer leakage.
2. **Layer Quality**: whether individual layers preserve clean boundaries, alpha masks, colors, textures, structures, and geometry without halos, aliasing, or residual artifacts.
3. **Occlusion Completion**: whether hidden regions are completed with plausible structures and remain consistent with visible content, without conspicuous hallucinations or artifacts.

4. **Recomposition Consistency**: whether alpha-compositing the ordered layers faithfully reconstructs the input image without gaps, duplication, opacity errors, incorrect occlusion relationships, or global distortion.
5. **Overall Quality**: the overall accuracy, visual quality, and practical usability of the decomposition as an editable layer representation.

The annotation interface is shown in Figure 6. For each sample, annotators are presented with the original image, two anonymized candidate layer decompositions, denoted as A and B, and their corresponding recompositions. Annotators compare the two candidates independently along the five dimensions defined above and select *A better*, *B better*, or *Tie* for each dimension.

A.6.1 Annotation Aggregation and Agreement

Each candidate pair is independently annotated by three annotators along all five evaluation dimensions. Before annotation, all annotators receive the same annotation guidelines, evaluation criteria, and representative examples to ensure a consistent interpretation of the task.

We adopt a conservative consensus rule to determine the final preference label for each dimension. We retain an annotation when either (i) all three annotators select the same option, or (ii) two annotators select the same directional preference (*A better* or *B better*) and the remaining annotator selects *Tie*. Cases with substantial disagreement are discarded, including those where all three annotators select different options, or where one annotator expresses a directional preference directly opposite to that of the other two. All other configurations without a clear non-conflicting preference consensus are also excluded from the final preference dataset.

To assess annotation consistency, we report the *unanimous agreement rate*, defined as the percentage of comparisons for which all three annotators select exactly the same option. As shown in Table 7, unanimous agreement ranges from 76.7% to 86.7% across the five dimensions, indicating a high level of consistency among annotators under the shared annotation protocol.

Using annotations collected from a pool of 20 computer-vision experts and the consensus-based filtering described above, we retain 13,979 high-quality pairwise preference annotations. We use 12,979 pairs for training and reserve 1,000 pairs for testing.

A.6.2 Annotation Examples

Figure 7 shows representative annotation examples. We provide the detailed rationale for each annotation below.

Example 1.

- **Decomposition Correctness (B better).** B keeps the peach-colored decorative shape and the "50K" text intact, while A splits the peach shape across Layers 1 and 2.
- **Layer Quality (B better).** B cleanly completes the peach-colored shape behind the text, while A leaves shadow-like residuals from the overlaid text on the completed region.
- **Occlusion Completion (B better).** B cleanly completes the occluded regions of the peach-colored shape, while A leaves shadow-like residuals from the overlaid text on the completed regions.
- **Recomposition Consistency (B better).** B faithfully reconstructs the peach-colored shape after recomposition, while A fails to reconstruct it seamlessly, resulting in visible cracks.
- **Overall Quality (B better).** B provides a cleaner decomposition of the woman, text, and decorative graphics with fewer fragmented elements.

Example 2.

- **Decomposition Correctness (A better).** A keeps "Happy Women's Day" and the floral decorations together, while B splits the title across three different layers.
- **Layer Quality (A better).** A preserves cleaner object boundaries, while B introduces thin contour artifacts along the edges of the woman and several decorative elements.

Table 7: Inter-annotator agreement across the five LAYERREWARD annotation dimensions. *Unanimous* denotes the percentage of comparisons for which all three annotators select exactly the same option.

Dimension	Unanimous
Decomposition Correctness	76.7%
Layer Quality	85.0%
Occlusion Completion	82.4%
Recomposition Consistency	78.6%
Overall Quality	86.7%

- **Occlusion Completion (B better).** B produces a cleaner white background after removing the central woman, while A leaves faint blurred residuals in Layer 0.
- **Recomposition Consistency (A better).** A more faithfully reconstructs the original image, whereas B retains visible thin outlines around foreground objects after recomposition.
- **Overall Quality (A better).** A provides a more coherent decomposition of the title and decorative elements, with fewer unnecessarily fragmented components.

Example 3.

- **Decomposition Correctness (B better).** B keeps "WORLD CANCER Day" together, while A separates "Day" from "WORLD CANCER" across different layers.
- **Layer Quality (Tie).** Both candidates preserve the silver top shape, blue ribbon, hands, and text with comparable visual quality.
- **Occlusion Completion (B better).** B avoids the detached gray patch appearing below "WORLD CANCER" in A's Layer 2.
- **Recomposition Consistency (Tie).** Both recompositions closely reproduce the pink background, top silver shape, blue ribbon, hands, and bottom decoration.
- **Overall Quality (B better).** B provides a more coherent grouping of the "WORLD CANCER Day" design with fewer fragmented or isolated components.

A.6.3 PREFERENCE ANALYSIS ACROSS DIMENSIONS

As shown in Figure 8(a), Decomposition Correctness and Overall Quality are the most discriminative dimensions, with 86.2% and 86.0% of comparisons receiving non-tie judgments, respectively. In contrast, Recomposition Consistency exhibits a higher tie rate. This is largely because many existing methods already achieve comparably strong performance in reconstructing the input from their decomposed layers, making differences along this dimension less visually apparent. Consistent with this Dobservation, our annotators reported that, for many samples, it was difficult to distinguish the recomposition quality of different models. Nevertheless, the large scale of our annotation effort still provides a substantial number of discriminative samples for training and evaluating the reward model.

Figure 8(b) presents the Pearson correlation heatmap among the five evaluation dimensions. Overall Quality is most strongly correlated with Decomposition Correctness ($r = 0.924$), followed by Layer Quality ($r = 0.801$) and Occlusion Completion ($r = 0.745$). Layer Quality and Occlusion Completion also exhibit a strong correlation ($r = 0.822$), which is intuitive because inaccurate occlusion completion can introduce visible artifacts that directly degrade the quality of the recovered layers, as illustrated by Example 1 in Figure 8. In comparison, Recomposition Consistency shows weaker correlations with the other dimensions. A likely reason is that many existing methods exhibit relatively small differences in recomposition quality, making this dimension less influential on the overall preference than dimensions that more directly reflect decomposition correctness and the perceptual quality of individual layers.

Table 8: Backbone ablation for LayerReward. Accuracy (%) is reported along the same five dimensions as in Table 9. Bold indicates the best result in each column, including ties.

Backbone	Decomp.	Layer Qual.	Occl.	Recomp.	Overall
Qwen3-VL-32B-Instruct	**83.80**	86.53	89.38	75.00	**84.83**
Qwen3-VL-8B-Instruct	83.69	**88.18**	**90.61**	74.61	84.24
Qwen3-VL-4B-Instruct	83.33	86.04	89.18	**75.39**	83.39
Qwen3-VL-2B-Instruct	81.31	86.54	88.57	**75.39**	82.07
Qwen2.5-VL-32B-Instruct	82.26	85.87	88.77	75.00	83.51
Qwen2.5-VL-7B-Instruct	82.03	85.88	87.14	75.00	82.31
Qwen2.5-VL-3B-Instruct	80.48	87.36	89.80	**75.39**	81.47

A.7 Ablation Study on LayerReward

We investigate two factors in LayerReward training: the amount of annotated preference data and the choice of backbone. We report pairwise preference classification accuracy along the five annotation dimensions defined in Appendix A.6.

A.7.1 Effect of Backbone Choice

Table 8 compares seven backbones from the Qwen3 (Bai et al., 2025a) and Qwen2.5 (Bai et al., 2025b) families. Within Qwen3, overall-quality accuracy increases with model size, from 82.07% for 2B to 84.83% for 32B. However, scaling does not uniformly improve every dimension. Qwen3-VL-32B-Instruct achieves the highest decomposition-correctness and overall-quality accuracies, whereas Qwen3-VL-8B-Instruct performs best on layer quality and occlusion completion, reaching 88.18% and 90.61%, respectively. Recomposition-consistency accuracy varies relatively little across the tested backbones, ranging from 74.61% to 75.39%.

Comparing backbone families, Qwen3-VL-32B-Instruct outperforms Qwen2.5-VL-32B-Instruct on four dimensions and matches it on recomposition consistency. At a smaller scale, Qwen3-VL-8B-Instruct exceeds Qwen2.5-VL-7B-Instruct on four dimensions, including gains of 1.93 percentage points in overall quality and 3.47 percentage points in occlusion completion, while trailing by 0.39 percentage points in recomposition consistency. Thus, backbone choice affects different quality criteria differently, and parameter count alone does not determine performance.

We select Qwen3-VL-8B-Instruct as the default LayerReward backbone. Its overall-quality accuracy is only 0.59 percentage points below that of Qwen3-VL-32B-Instruct, while it achieves the strongest layer-quality and occlusion-completion results with one quarter of the nominal parameter count. This balance motivates its use both as a benchmark evaluator and as a repeatedly queried reward model during preference-aligned post-training, without requiring the larger 32B backbone.

A.7.2 Effect of Preference Data Scale

We train reward models with 20%, 50%, and 100% of the preference training set using Qwen3-VL-8B-Instruct and Qwen2.5-VL-7B-Instruct backbones. Table 9 shows that increasing the training fraction consistently improves accuracy across all five dimensions for both backbones. For Qwen3-VL-8B-Instruct, increasing the fraction from 20% to 100% raises overall-quality accuracy from 80.63% to 84.24% and occlusion-completion accuracy from 85.10% to 90.61%. Recomposition consistency exhibits the largest gain, increasing from 62.50% to 74.61%, or 12.11 percentage points. Qwen2.5-VL-7B-Instruct follows the same trend: overall-quality accuracy rises from 79.66% to 82.31%, while recomposition-consistency accuracy improves from 66.80% to 75.00%.

The gains persist when moving from 50% to the full training set. In particular, Qwen3-VL-8B-Instruct improves overall-quality accuracy by 2.29 percentage points and occlusion-completion accuracy by 2.85 percentage points. These results support the value of additional task-specific preference supervision for both perceptual assessment and compositional reasoning. The consistent improvements across both backbones further suggest that the benefits of the annotations are not restricted to a single backbone family.

Table 9: Effect of preference training-set size on LayerReward. We report accuracy (%) for decomposition correctness (Decomp.), layer quality (Layer Qual.), occlusion completion (Occl.), recomposition consistency (Recomp.), and overall quality (Overall). Higher is better; bold indicates the best result within each backbone.

Backbone	Training set	Decomp.	Layer Qual.	Occl.	Recomp.	Overall
Qwen3-VL-8B-Instruct	20%	79.52	84.24	85.10	62.50	80.63
	50%	81.43	85.88	87.76	72.66	81.95
	100%	**83.69**	**88.18**	**90.61**	**74.61**	**84.24**
Qwen2.5-VL-7B-Instruct	20%	78.10	79.47	84.90	66.80	79.66
	50%	80.12	84.24	86.33	71.09	81.59
	100%	**82.03**	**85.88**	**87.14**	**75.00**	**82.31**

Table 10: Hyperparameter configurations for LAYERCOMPASS post-training.

Hyperparameter	**Value**
Training hardware	2×8 NVIDIA A800 (80GB)
Policy / Reward GPU split per node	7 / 1
Rollout group size (K)	7
Training steps	200
Max pixel budget	640×640 (native aspect ratio)
Base model	QWEN-IMAGE-LAYERED
Optimizer	AdamW
Learning rate	2×10^{-4}
LoRA rank (r)	64
LoRA alpha (α)	64
LoRA dropout	0.0
NFT factor(β)	0.1
NFT decay type	1

A.8 TRAINING DETAILS OF LAYERCOMPASS

In this section, we elaborate on the implementation and training configurations of LAYERCOMPASS. The preference-aligned post-training is conducted across two compute nodes equipped with a total of $16\times$ NVIDIA A800 GPUs. Within each 8-GPU node, 7 GPUs are allocated for policy rollout generation and backward parameter updates, while the remaining 1 GPU is dedicated to hosting the LAYERREWARD model for real-time online preference scoring. Under this distributed setup, the rollout group size is set to $K = 7$ candidates per prompt, corresponding to the parallel rollout generation capacity of each node.

LAYERCOMPASS is fine-tuned upon QWEN-IMAGE-LAYERED using LoRA on the diffusion transformer backbone while keeping the RGBA-aware VAE frozen. We optimize the model using the DIFFUSIONNFT objective (Section 3.5) for 200 optimization steps. The detailed hyperparameter specifications are summarized in Table 10.

The training and evaluation curve of LAYERCOMPASS is shown in Figure 9 and 10

A.9 HUMAN PREFERENCE STUDY

To directly validate the LayerBench ranking, we conduct a blind A/B study between LayerCompass and its base model, Qwen-Image-Layered. We sample 200 prediction pairs uniformly from the six evaluation subsets, randomize the display sides, and hide all model identities. Three independent raters judge which decomposition is better overall (with an explicit tie option), inspecting both the recomposited result and the individual layers of each prediction.



Figure 9: Training reward curve of layercompass



Figure 10: Evaluation reward curve of layercompass

Table 11 summarizes the results. LayerCompass is preferred on 76.4% of the 600 votes (95% CI [72.0%, 80.6%]), well above chance, and wins the strict majority vote on 79.0% of pairs. Rater consistency is high: the three raters' decisive preference rates differ by less than 1.2 percentage points (0.800, 0.798, 0.788), they agree unanimously on 60% of pairs and two-out-of-three on a further 35.5%; only 12 pairs (6%) lack a strict non-tie majority and are excluded from the majority-vote statistic. The per-dataset breakdown (Figure 11) shows that the preference is consistent across domains but varies in magnitude: LayerCompass dominates on Editability (90.9% of votes) and PSD (80.4%), while Photo is the most contested domain—although LayerCompass still wins more decisive votes there (38.2% vs. 23.5%), raters cannot distinguish the two models on 38.2% of Photo pairs, indicating that photorealistic decomposition quality is the most perceptually ambiguous. These human preferences corroborate the LayerBench rankings, where LayerCompass leads Qwen-Image-Layered on the final score (0.703 vs. 0.592).

A.10 Cross-Model Human Ranking Study

To further examine whether the preference-aligned evaluation reflects human judgment across different decomposition paradigms, we conduct a five-way blind ranking study over all evaluated models. We randomly sample 40 LayerBench images and collect rankings from four independent raters for each image. For every sample, raters jointly inspect the five anonymized layer decompositions and their recompositions, with model positions independently randomized for each rater. Ties are explicitly allowed to avoid forcing artificial distinctions between perceptually similar outputs.

Table 13 reports the resulting human preferences. LayerCompass obtains the best mean rank (1.65), the highest Rank-1 frequency (57.5%), and the highest pairwise win share (69.7%). Stable-Layers ranks second, followed by Qwen-Image-Layered, SynLayers, and LayerD. The ranking is consistent across raters, with a mean pairwise Kendall τ_b of 0.69 and 88.3% agreement on the top-ranked model.

Notably, the resulting human ordering,

$$\text{LayerCompass} > \text{Stable-Layers} > \text{Qwen-Image-Layered} > \text{SynLayers} > \text{LayerD},$$

exactly matches the ordering produced by the LayerReward preference score. This provides an independent cross-model validation of LayerReward beyond its held-out pairwise test set, and suggests that it captures aspects of holistic decomposition quality that are not fully reflected by property-specific metrics. In particular, SynLayers performs strongly on text preservation and operational editability, yet receives lower holistic human preference, illustrating that strong performance on individual editing probes does not necessarily imply better overall layer organization.

This study is intended as an independent validation of the human-preference-aligned component rather than as a replacement for the benchmark's deterministic metrics. The final LayerBench score

Table 11: Blind A/B human preference between LayerCompass and Qwen-Image-Layered. Three raters each compare 200 prediction pairs sampled uniformly from the six evaluation subsets (600 votes in total; sides are randomized and model identities are hidden). Pref. share counts a tie as half a vote for each side; the confidence interval is a bootstrap 95% CI resampled over pairs. Majority vote requires a strict non-tie 2-of-3 majority.

LC Win	Tie	QW Win	Pref. Share	95% CI	Majority
427 (71.2%)	63 (10.5%)	110 (18.3%)	0.764	[0.720, 0.806]	158/200 (79.0%)

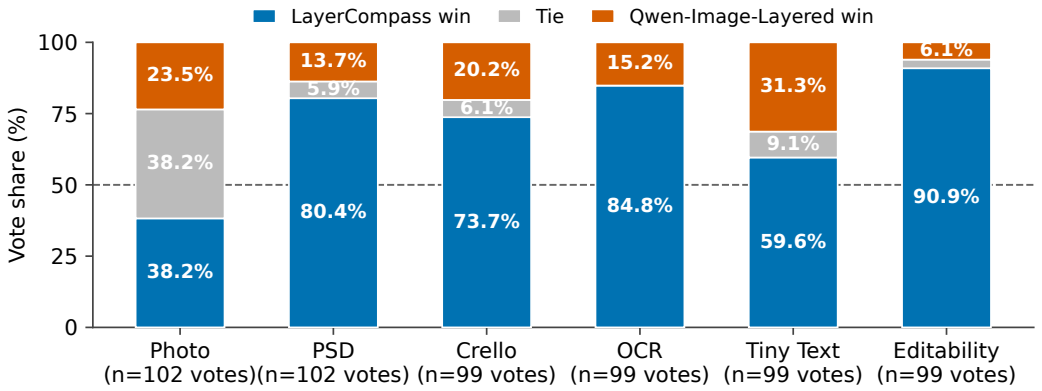


Figure 11: Per-dataset vote distribution in the blind A/B study. Each bar aggregates the three raters' votes; the dashed line marks the 50% chance level.

therefore retains both sources of evidence: benchmark-designed metrics capture explicit and reproducible properties, while LayerReward captures holistic human preference. Their complementary behavior motivates the balanced aggregation used in the main benchmark.

A.11 Sensitivity to Aggregate-Score Weighting

The LayerBench final score balances benchmark-designed evaluation (rule-based and VLM-based metrics) with human-preference-aligned evaluation. To examine whether the aggregate ranking depends critically on the specific weighting, we vary the LayerReward weight w_{LR} while equally splitting the remaining weight between Rule and VLM:

$$S(w_{\mathrm{LR}}) = \frac{1 - w_{\mathrm{LR}}}{2} S_{\mathrm{Rule}} + \frac{1 - w_{\mathrm{LR}}}{2} S_{\mathrm{VLM}} + w_{\mathrm{LR}} S_{\mathrm{LR}}. \tag{34}$$

As shown in Table 12, LayerCompass remains the top-ranked method for $w_{\mathrm{LR}} \in [0.4, 0.6]$, including the default balanced setting $w_{\mathrm{LR}} = 0.5$. At the near-equal three-family weighting ($w_{\mathrm{LR}} = 0.33$), SynLayers ranks first due to its particularly strong rule-based and VLM editability scores, while LayerCompass remains second.

This behavior is consistent with the intended interpretation of the composite score rather than indicating a single universally correct weighting: benchmark-designed metrics emphasize explicit measurable properties, whereas LayerReward captures holistic human preference. Importantly, the independent five-model human ranking study yields the same model ordering as LayerReward, providing external support for retaining substantial preference-aligned weight in the aggregate score. Overall, LayerCompass's leading result does not depend on the exact default weight of 0.5, but holds across a broad range of preference-aligned weightings.

A.12 Design Choice for Evaluator Deployment.

For practical benchmark deployment, we intentionally use the same Qwen3-VL-8B-Instruct backbone for the generic VLM judge and LayerReward. This avoids requiring users to host multiple

Table 12: **Sensitivity to aggregate-score weighting.** The remaining weight $1 - w_{\mathrm{LR}}$ is equally divided between Rule and VLM. The main benchmark setting uses $w_{\mathrm{LR}} = 0.5$. **Bold**: best; underline: second best.

Model	w_{LR}=0.33	w_{LR}=0.40	w_{LR}=0.50	w_{LR}=0.60
LayerD	0.473	0.475	0.480	0.484
SynLayers	**0.690**	0.683	0.674	0.665
Stable-Layers	0.601	0.615	0.636	0.656
Qwen-Image-Layered	0.559	0.572	0.592	0.612
LayerCompass	0.679	**0.689**	**0.703**	**0.718**

Table 13: **Human preference study.** Each of 4 raters ranks the layer decompositions produced by all five models on the same image (40 samples, ties allowed; candidate positions are randomized per rater). *Mean Rank* is averaged over all 160 rankings (lower is better). *Rank-1* is the fraction of rankings placing a model first (shared first places included). *Win Share* treats ties as 0.5; *Decisive* considers only pairwise comparisons without ties. Inter-rater agreement: mean pairwise Kendall $\tau_b = 0.69$, rank-1 agreement 88.3%.

Model	Mean Rank ↓	Rank-1 (%) ↑	Win Share (%) ↑	Decisive (%) ↑
LayerCompass	**1.65**	**57.5**	**69.7**	**76.2**
Stable-Layers	1.99	48.7	60.3	63.4
Qwen-Image-Layered	2.35	42.5	50.0	50.0
SynLayers	2.39	28.7	42.2	41.1
LayerD	2.86	7.5	27.8	26.2

large VLM backbones during evaluation. The two evaluators nevertheless differ substantially in how they are instantiated: the VLM judge is prompt-based and evaluates explicit editing probes, whereas LayerReward is preference-tuned with LoRA and a task-specific five-dimensional reward head using expert pairwise annotations.

A.13 Qualitative Diagnosis of Metric-Specific Failures

We visualize representative cases in which different methods perform poorly on specific evaluation dimensions. These examples provide an intuitive interpretation of what each metric captures.

Reconstruction fidelity. As shown in Figure 12, Stable-Layers exhibits noticeable color shifts between the input and recomposed images, including changes in texture and overall color tone. These appearance deviations lead to low reconstruction scores of 0.115–0.211 in the shown examples.

Layer organization. As shown in Figure 13, LayerD often fragments coherent visual elements across multiple layers, leading to duplicated content, residual artifacts, and foreground–background leakage. These structural inconsistencies result in relatively low layer-alignment scores of 0.499–0.641.

Structural decomposition failures. Figure 14 illustrates both successful and failed decompositions produced by SynLayers. For relatively simple examples, the predicted layers closely match the underlying visual structure, achieving alignment scores of 0.961 and 0.912 with reconstruction scores of 0.865 and 0.946. In the second example, however, a coherent object is excessively fragmented into multiple small layers, reducing the alignment score to 0.348 and the reconstruction score to 0.511. The fourth example exhibits a more severe prediction failure, where the generated layers no longer correspond correctly to the visual content of the input, resulting in an alignment score of 0.355 and a reconstruction score of only 0.197. These examples show that low alignment scores can arise from qualitatively different structural errors, ranging from over-fragmentation to incorrect layer prediction.

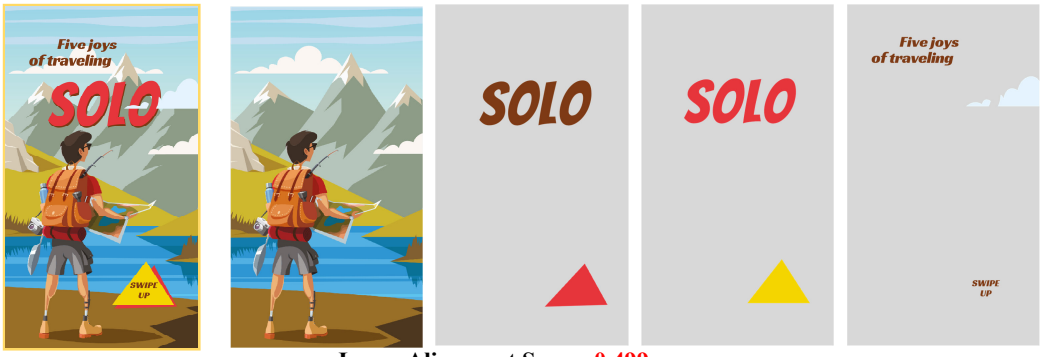
Variation across evaluation dimensions. Figure 15 further illustrates that decomposition quality can vary substantially across samples and evaluation dimensions. Qwen-Image-Layered achieves



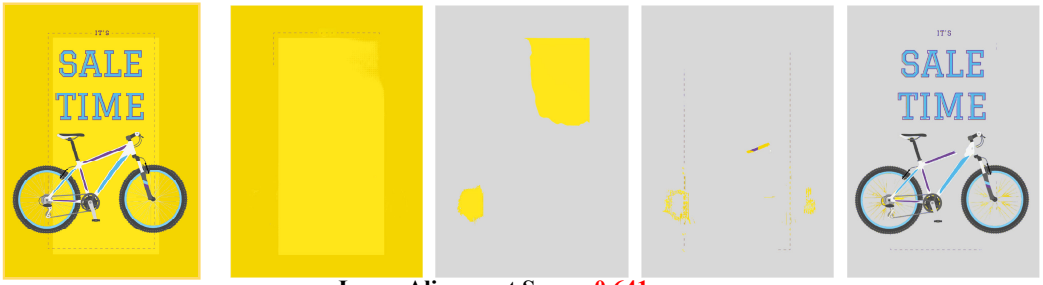
Figure 12: Representative reconstruction failures of Stable-Layers. Noticeable color shifts between the input and recomposed images lead to low reconstruction scores.

strong layer alignment (0.914) and reconstruction (0.928) on a successful example, whereas more challenging layouts obtain lower alignment scores of 0.567 and 0.665 and reconstruction scores of 0.559 and 0.562. Meanwhile, the text-rich example achieves a text-preservation score of 0.914. These cases highlight that reconstruction, structural decomposition, and content preservation capture complementary aspects of decomposition quality.

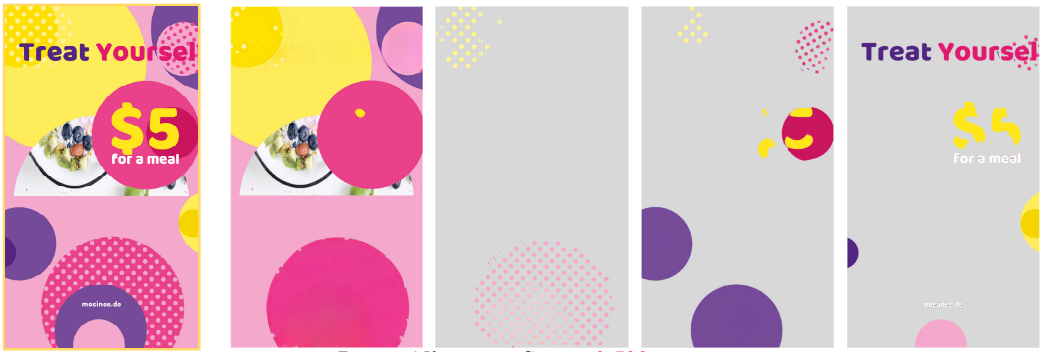
Taken together, these visualizations provide qualitative evidence for the diagnostic role of the proposed metrics. Reconstruction captures discrepancies in the final composite, layer alignment reflects structural correspondence between decompositions, and text preservation evaluates the fidelity of textual content. The distinct failure modes observed across these examples further motivate the multi-dimensional evaluation protocol adopted in LayerBench.



Layer Alignment Score: 0.499



Layer Alignment Score: 0.641



Layer Alignment Score: 0.500



Layer Alignment Score: 0.527

Figure 13: Representative layer-alignment failures of LayerD. Coherent visual elements are fragmented across multiple layers, with duplicated content, residual artifacts, and foreground–background leakage.

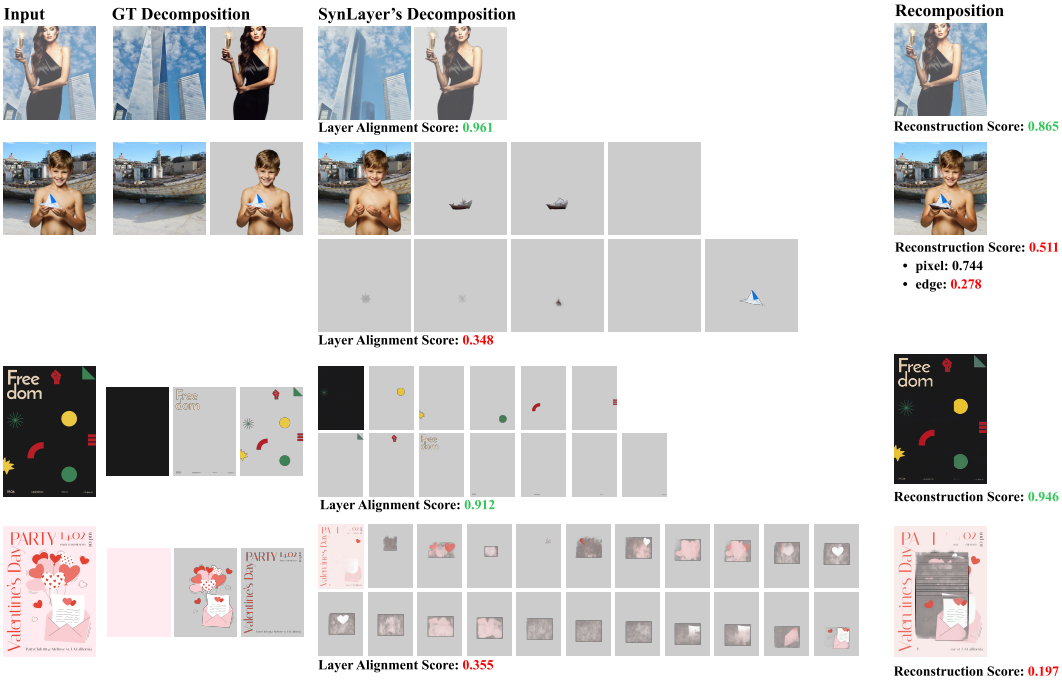


Figure 14: Representative SynLayers results showing distinct structural failures, including over-fragmentation and incorrect layer prediction.

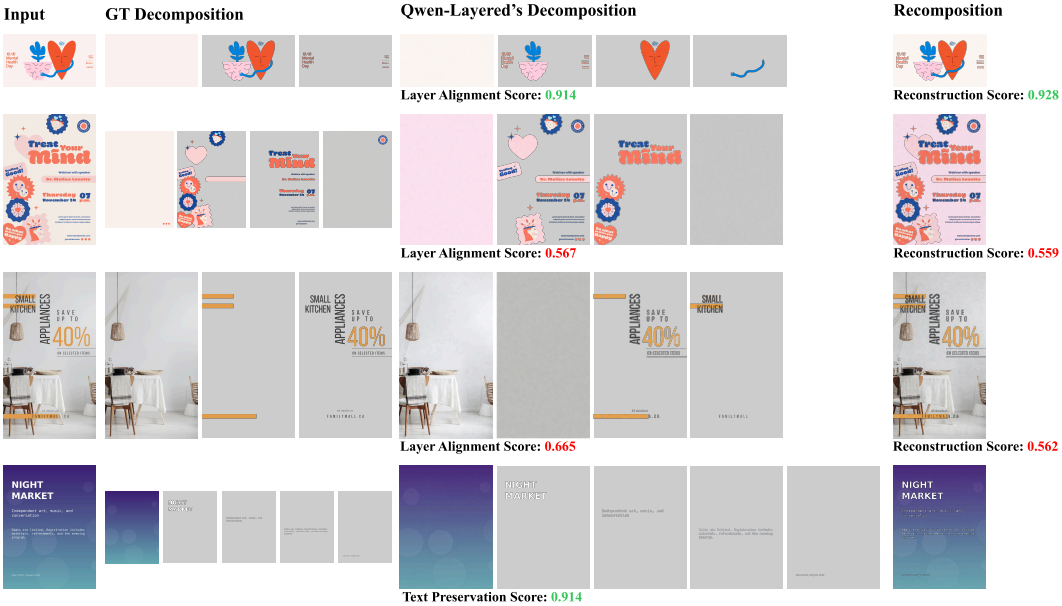


Figure 15: Representative Qwen-Image-Layered results across evaluation dimensions, illustrating variation in layer alignment, reconstruction fidelity, and text preservation.