PosterOmni: Generalized Artistic Poster Creation via Task Distillation and Unified Reward Feedback

Supplementary Material

This is supplementary material for *PosterOmni: Generalized Artistic Poster Creation via Task Distillation and Unified Reward Feedback*.

We present the following materials in this supplementary document:

- **Sec. 1** Prerequisites for flow matching and reinforcement learning, including background on velocity parameterization, rectified flows, GRPO, and DiffusionNFT.
- **Sec. 2** Details of our PosterOmni data suite (PosterOmni-200K and PosterOmni-Bench), covering prompt design, multimodal filtering, task-specific image-to-poster construction pipelines, and keyword/topic coverage.
- **Sec. 3** Construction of the PosterOmni reward training dataset and implementation details of the unified PosterOmni reward model R_{omni}.
- **Sec. 4** User study setup and results, including the human evaluation protocol and win/tie/loss statistics against open-source and proprietary baselines.
- **Sec. 5** Additional ablation studies on reward model design and expert integration strategies.
- **Sec. 6** Additional visual comparisons across all six image-to-poster tasks, illustrating qualitative differences between PosterOmni and competing methods.
- **Sec. 7** Limitations and future work of the PosterOmni.

1. Prerequisites for Flow Matching and Reinforce Learning

Flow Matching and Velocity Parameterization. Diffusion models [6, 18] generate samples by reversing a forward noising process, which can be written as a deterministic trajectory

$$x_t = \alpha_t x_0 + \sigma_t \epsilon, \quad \epsilon \sim \mathcal{N}(0, I),\ t \in [0, 1], \tag{1}$$

where α_t and σ_t describe the evolution of the signal and noise, respectively. The velocity parameterization [24] predicts the tangent of this diffusion trajectory. Let

$$v = \dot{\alpha}_t x_0 + \dot{\sigma}_t \epsilon \tag{2}$$

denote the instantaneous velocity along x_t. A neural network $v_\theta(x_t, t, c)$ is then trained to approximate this target field by minimizing

$$\mathbb{E}_{t,x_0,\epsilon}\big[w(t)\|v_\theta(x_t, t, c) - v\|_2^2\big], \tag{3}$$

where $w(t)$ is a time-dependent weight. Sampling is performed by solving the deterministic ODE of the forward process:

$$dx_t = v_\theta(x_t, t, c)\, dt. \tag{4}$$

Rectified flow [11, 13] can be viewed as a simplified instance of this velocity-parameterized formulation. Given a data sample $x_0 \sim X_0$ with condition c and a Gaussian sample $x_1 \sim X_1$, it constructs the linear interpolation

$$x_t = (1 - t)x_0 + t x_1, \quad t \in [0, 1], \tag{5}$$

whose velocity field satisfies $v = x_1 - x_0$. The corresponding flow-matching objective is

$$\mathcal{L}_{\text{FM}}(\theta) = \mathbb{E}_{t,x_0,x_1}\left[\|v - v_\theta(x_t, t, c)\|_2^2\right]. \tag{6}$$

This setting is recovered from the diffusion trajectory by choosing $\alpha_t = 1 - t$ and $\sigma_t = t$, which yields $v = \dot{\alpha}_t x_0 + \dot{\sigma}_t \epsilon = \epsilon - x_0$; identifying x_1 with ϵ recovers the rectified flow interpolation between x_0 and a Gaussian sample x_1.

Policy-Gradient Reinforce Learning for Diffusion Flows. Recent works [9, 12, 20, 22] formulate diffusion sampling as a multi-step Markov Decision Process (MDP), which enables the use of policy gradient methods such as PPO and GRPO. For rectified flows, however, the purely deterministic ODE dynamics prevent direct application of GRPO. FlowGRPO [12] addresses this issue by introducing stochasticity through an SDE under the velocity parameterization:

$$dx_t = \Big[v_\theta(x_t, t) + \frac{g_t^2}{2t}\big(x_t + (1 - t)v_\theta(x_t, t)\big)\Big]dt + g_t\, d\omega_t, \tag{7}$$

where

$$g_t = a\sqrt{\frac{t}{1 - t}} \tag{8}$$

controls the magnitude of injected noise and a is a tunable scale.

Discretizing this SDE with an Euler step of size Δt yields a Gaussian transition kernel between adjacent states:

$$\pi_\theta(x_{t-\Delta t} \mid x_t) = \mathcal{N}\Big(x_t + \Big[v_\theta(x_t, t) + \frac{g_t^2}{2t}\big(x_t + (1 - t)v_\theta(x_t, t)\big)\Big]\Delta t,\ g_t^2 \Delta t\, I\Big). \tag{9}$$

Such a parameterization makes the reverse-time transitions likelihood-tractable Gaussians, allowing existing policy gradient algorithms (e.g., GRPO) to be directly applied to diffusion models.

Figure 1. **Examples from our PosterOmni-data (PosterOmni-200K and PosterOmni-Bench).** For each of the six core image-to-poster tasks—style-driven generation, layout-driven generation, ID-driven generation, extending, rescaling, and filling—we show the reference image(s) together with the corresponding image-to-poster prompts in both English and Chinese. The examples illustrate diverse commercial scenarios, layouts, and visual styles, as well as the explicit task-specific instructions.

Diffusion Negative-aware Finetuning (DiffusionNFT). DiffusionNFT [25] performs direct policy optimization on the forward diffusion process by leveraging a reward signal $r(x_0, c) \in [0, 1]$. Rather than using standard policy gradient [12, 22], it forms a contrastive diffusion loss that pushes the model's velocity predictor toward high-reward behavior and away from low-reward behavior.

Given an offline diffusion policy v^{old}, DiffusionNFT constructs implicit positive and negative policies:

$$v_\theta^+(x_t, t, c) = (1-\beta)v^{\text{old}}(x_t, t, c) + \beta v_\theta(x_t, t, c), \quad (10)$$

$$v_\theta^-(x_t, t, c) = (1+\beta)v^{\text{old}}(x_t, t, c) - \beta v_\theta(x_t, t, c), \quad (11)$$

where β controls guidance strength. The training objective is

$$\mathcal{L}(\theta) = \mathbb{E}_{c,\pi^{\text{old}}(x_0|c),t}\left[r\left\|v_\theta^+ - v\right\|_2^2 + (1-r)\left\|v_\theta^- - v\right\|_2^2\right], \quad (12)$$

directly optimizing the new velocity field toward a reward-weighted improvement direction. The reward is normalized as:

$$r(x_0, c) = \frac{1}{2} + \frac{1}{2}\,\text{clip}\left(\frac{r^{\text{raw}}(x_0, c) - \mathbb{E}_{\pi^{\text{old}}} r^{\text{raw}}(x_0, c)/Z_c}{1}, -1, 1\right). \quad (13)$$

where Z_c normalizes global reward scale. Unlike policy-gradient diffusion RL, DiffusionNFT maintains forward consistency, integrates reinforcement signals implicitly into the velocity field, and entirely avoids likelihood approximation—enabling a simple, stable finetuning mechanism on the forward diffusion dynamics.

2. Details of PosterOmni Data (PosterOmni-200K and PosterOmni-Bench)

In this section, we provide additional details of our data suite PosterOmni data, which consists of the training set **PosterOmni-200K** and the evaluation benchmark **PosterOmni-Bench**. The main paper briefly introduces the automated pipeline in Sec. 3.1 of our manscript; here we elaborate on task-specific construction, multimodal filtering, and topic coverage.

2.1. Prompt Design and Base Text-to-Image Corpus

We first build a diverse text-to-image corpus that mimics real poster design scenarios. Following the meta-prompt in Fig. 12, we sample a *category* (e.g., products, food, events/travel, education, nature, entertainment), a *scenario* (e.g., “family feast”, “AI summit”), and a *style tag* (e.g., Swiss grid, watercolor). For each triplet, a VLM (GPT [14]/Qwen3 [23]) plays the role of a “creative director” and writes a fluent image-to-poster prompt specifying (1) main subjects, (2) spatial composition, (3) overall mood and color palette, and (4) 1–3 pieces of rendered text with approximate positions (title, slogan, time/place).

We instantiate the template in both English and Chinese, leading to bilingual prompts with consistent seman-

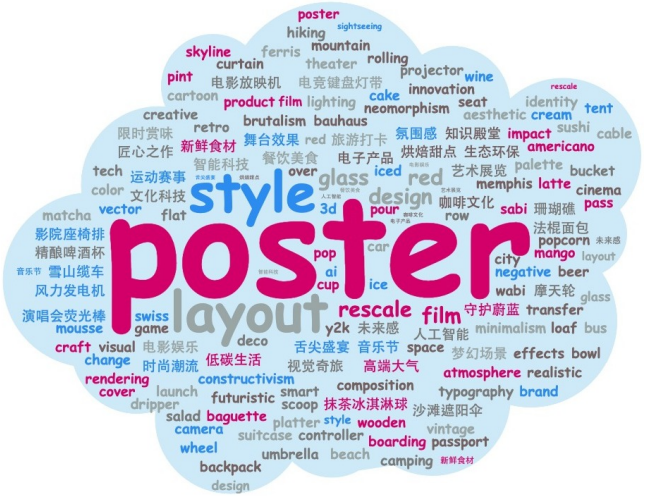

Figure 2. **Word cloud of representative keywords in PosterOmni-data.** It illustrates the diversity of our dataset in terms of styles, layouts, contents and themes, covering a broad range of real-world poster scenarios in both English and Chinese.

tics. Multiple candidate images are then generated per prompt by strong text-to-image models (Qwen-Image [21] and FLUX-style generators), which form the base images from which all downstream tasks are constructed.

2.2. Multimodal Filtering for PosterOmni-200K and PosterOmni-Bench

To ensure that the synthetic posters are usable for supervised training and reliable evaluation, we apply a multi-stage multimodal filtering pipeline (Fig. 2 in our manscript).

Training set (PosterOmni-200K). For each (prompt, image) pair we perform:

- **OCR and text sanity checks.** PaddleOCR [3] is used to extract rendered text; we reject images where the decoded strings are unreadable or deviate too much from the prompt keywords (e.g., wrong language, heavy corruption).
- **Vision–language consistency.** We embed both the prompt and image with Jina-clip-v2 [8] and drop samples whose similarity falls below a threshold, which removes cases where the layout or subject semantics are obviously mismatched with the description.
- **Layout and clutter constraints.** Simple heuristics on text box count, text area ratio, and foreground–background separation filter out extremely cluttered or almost text-free images, so that each sample still resembles a reasonable poster layout.

Only pairs passing all checks are used as sources for building task-specific image-to-poster pairs.

Benchmark (PosterOmni-Bench). For the test benchmark, we adopt stricter filtering:

- **Task suitability via VLM.** Given a candidate image, Gemini-2.5-Flash [19] is queried with the task-matching meta-prompt in Fig. 13. It must assign exactly one label from {`EXTENDING`, `FILLING`, `RESCALING`, `ID-DRIVEN`, `LAYOUT-DRIVEN`, `STYLE-DRIVEN`, `NONE`}. We keep only samples that receive a confident, non-`NONE` label consistent with our intended task.
- **Manual spot-checking.** For each task and theme, we manually review a subset of images to verify that the predicted task matches human intuition (e.g., that an "extending" candidate indeed has expandable background, that a "layout" case exposes a clear grid/compositional structure).

This procedure yields 480 English and 540 Chinese prompts paired with reference images, balanced across six themes and six tasks, as described in the main paper.

2.3. Task-Specific Image-to-Poster Construction

Starting from the filtered text-to-image corpus, we construct paired image-to-poster samples for six tasks by applying modular, task-specific transformations (also summarized in Fig. 2 of our manscript). Examples of the resulting input–output pairs are shown in Fig. 1.

Extending. Given a reference poster, SAM-2 [15] is used to segment the main subject and foreground elements. We dilate the foreground mask and treat the remaining area as "extendable" background. The input image is obtained by cropping the canvas around the subject, while the target poster retains the original full canvas. The image-to-poster prompt explicitly asks the model to extend the canvas, this encourages learning seamless background extension and composition completion.

Filling. For the filling task, we first sample one or more localized regions (such as a logo slot, product placeholder, or empty billboard) using SAM-2 [15] masks and simple geometric rules. The input is created by erasing these regions to obtain a hole image; the target is the original poster. Fill prompts ask the model to regenerate appropriate content inside the hole (e.g., "replace the empty stand with a perfume bottle") under the same style and lighting.

Rescaling. Rescaling pairs simulate aspect-ratio changes without distorting the main subjects. For a clean poster, we use the SOTA commercial models/crop methods to change the central region to a different aspect ratio (e.g., from $2:3$ to $4:3$) and use BrushNet [7] to extend the margins where necessary so that the central content stays intact. The cropped or partially extended image is used as the input, while the fully adjusted poster serves as the output. Prompts describe the target ratio (as in PosterOmni-Bench) to teach the model aspect-ratio-aware composition.

Identity-driven generation. We treat the original poster as an ID reference and generate a new scene featuring the same key subject. PaddleDet [1] first detects identity-critical objects (e.g., a specific drink can, branded product, or mascot). We then use SOTA models to synthesize new im-

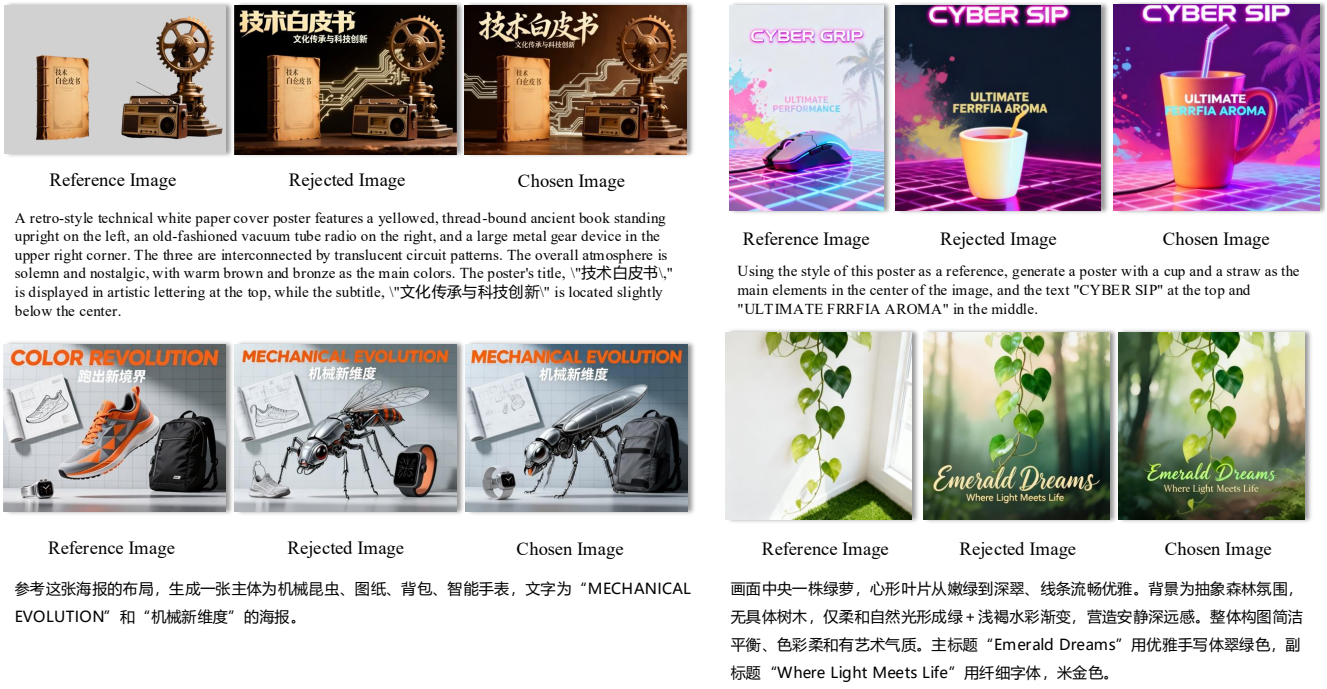

Figure 3. **Examples of preference pairs for PosterOmni Reward Training.** For several representative style-driven and layout-driven cases, we show the reference image together with the rejected and chosen candidates produced by PosterOmni-SFT, as well as the corresponding image-to-poster prompts in English and Chinese. Each triplet (reference, rejected, chosen) constitutes a concrete example of the preference pairs used to train the unified reward model R_{omni}.

age(s) where the subject appears in a different pose or environment but with consistent fine-grained identity (shape, color pattern, logo). The input consists of the generated image(s), and the output is the reference poster, supervised by prompts that stress preserving identity while changing context.

Layout-driven generation. Here the input is a clean layout template with recognizable blocks (hero image area, text zones, logo strip, etc.). We use VLMs (e.g., Gemini-2.5-Pro [19]) or simple heuristic rules to extract a coarse layout graph, then ask the SOTA models to “follow the layout” but replace the content (e.g., new products and background). Then, SOTA VLM feedback is used to construct image-to-poster prompts to form a complete data pair.

Style-driven generation. For style-driven generation, the construction is analogous but focuses on visual treatment rather than spatial structure. Given a reference poster, we treat it as a style template and use VLMs to summarize key stylistic attributes such as color palette, rendering texture, lighting, and typography (e.g., “vaporwave cyberpunk”). We then require the existing editing model to replace parts of the text and objects in the scene while preserving reasonable consistency with the main stylistic features and scene semantics. The VLM is then used again to generate corresponding image-to-poster conversion prompts. Therefore, the input reference poster and target poster are stylistically similar but differ in specific content.

2.4. Keyword Distribution and Topic Coverage

To better visualize the semantic coverage of PosterOmni-data, Fig. 2 shows a word cloud built from all English and Chinese prompts. Large keywords such as “poster”, “layout”, “style”, “rescale”, and “film” correspond to our core tasks and typical poster scenarios, while medium and small words cover product categories (e.g., coffee, skincare, camera), event types (e.g., concert, marathon, exhibition), and design attributes (e.g., “minimalism”, “memphis”, “cream tone”). The mixture of bilingual tokens indicates that the dataset spans both Chinese and English markets and emphasizes realistic commercial usage rather than toy scenes. In addition to the high-quality data generated by our pipeline, PosterOmni-data also includes a small portion ($< 10\%$) of in-house poster data; these samples are processed with the same processes.

3. PosterOmni Reward Training Dataset and Model Details

3.1. PosterOmni Reward Dataset Construction

To clarify the data used for training the unified PosterOmni Reward Model R_{omni}, we summarize the construction pipeline and basic statistics here. As illustrated in

Table 1. Approximate statistics of the PosterOmni preference dataset used to train R_{omni}. Each row reports the number of human-checked preference pairs for a task type. During reward training, we further augment the data with input–output negative pairs $(I_{\text{in}}, I_{\text{chosen}})$; the last column shows the approximate fraction of such extra negatives among all comparisons.

Task type	#Preference pairs	Share of all pairs	Extra negative-pair ratio
Poster Rescale	11,000	$\approx 18\%$	33.3% (1:3)
Poster Fill	9,000	$\approx 15\%$	33.3% (1:3)
Poster Extend	10,000	$\approx 17\%$	33.3% (1:3)
Identity-driven	8,000	$\approx 13\%$	33.3% (1:3)
Layout-driven	11,000	$\approx 18\%$	33.3% (1:3)
Style-driven	11,000	$\approx 18\%$	33.3% (1:3)
Overall	60,000	100%	33.3% (1:3)

Fig. 3, starting from the SFT-trained PosterOmni model, we generate candidate posters for all six image-to-poster task types. Candidate images are grouped into pairs, each pair sharing the same input context and task description. We then query Gemini-2.5-Pro [19] with the preference prompt shown in Fig. 14 to obtain an automatic choice between the two candidates. Pairs for which Gemini indicates a clear preference and at least one candidate already satisfies basic poster quality are kept, while pairs where both candidates are obviously broken (e.g., unreadable text, collapsed layout) are discarded. This step acts as a coarse filter and provides an initial ranking signal.

On the remaining pairs, human annotators perform a light review using the same task-specific criteria as in Fig. 14, correcting Gemini's decisions when necessary and discarding ambiguous or noisy cases. After this two-stage filtering and review, we obtain roughly 60K clean preference pairs across all tasks. The distribution over task types is slightly imbalanced but covers both local editing (rescale, fill, extend, ID-driven) and global creation (layout-driven, style-driven) cases. For reward training, each labeled pair $(I_{\text{chosen}}, I_{\text{rejected}})$ additionally yields a simple negative pair by treating the original input image I_{in} as the less preferred sample and I_{chosen} as the preferred one, so that R_{omni} learns to favor complete poster-like edits over raw inputs. Tab. 1 reports the approximate per-task statistics used in our experiments.

3.2. PosterOmni Reward Model Architecture and Training

Based on the preference data described above, we instantiate R_{omni} on top of the Qwen3VL [23] encoder with a lightweight regression head. For each quadruplet $(I_{\text{in}}, p_t, \text{edit}, I)$, we treat I as the candidate poster to be scored. The image I is fed to the vision branch of Qwen3VL, while the text branch concatenates the task prompt p_t, the editing description `edit`, and a short task-type tag (e.g., "[Task: Layout-driven generation]"). We take the pooled multimodal representation and pass it through a small MLP head to obtain a scalar reward $r_\theta(I) \in \mathbb{R}$. Since each task is accompanied by explicit instructions and a task-type indicator, the reward model learns to distinguish fine-grained quality differences between candidates under the *same* task while sharing parameters across *different* tasks. In this way, preference learning mainly depends on relative scores within each task, yet results in a single unified PosterOmni reward model applicable to all image-to-poster settings.

4. User Study

Besides the automatic metrics reported the in mainscript, we further conduct a human preference study to directly assess the perceptual quality of different image-to-poster generation systems. Our goal is to measure how often PosterOmni is preferred by human users compared with both open-source and proprietary baselines.

Setup. We randomly sample 150 prompts from PosterOmni-Bench-en (in order to compare all models), covering all six poster-editing tasks (extend, fill, rescale, ID-driven, layout-driven, and style-driven generation). For each prompt, we generate posters using PosterOmni and six competing systems (Seedream-4.0 [16], Seedream-3.0 [5], UniWorld-V2–Qwen-Image-Edit [10], Qwen-Image-Edit [2509] [21], FLUX.1 Kontext [dev] [2], and BAGEL [4]). We recruit six experienced poster designers, all of whom have at least two years of professional design experience. Each rater is presented with pairwise comparisons between PosterOmni and one baseline at a time, under a randomized order of prompts and model sides (left-/right) to avoid bias.

Protocol and metrics. For each comparison, raters are asked to judge the two posters along four criteria: (i) *Aesthetic Value* (overall visual appeal and layout harmony), (ii) *Task (Prompt) Alignment* (whether the poster correctly follows the editing instruction and preserves required content/layout), (iii) *Text Accuracy* (correctness and legibility of rendered text), and (iv) *Overall Preference* (which poster they would choose to use in a real project). For every criterion, raters choose one of three options: "PosterOmni is better", "Tie", or "Baseline is better". Given all annotations, we compute for each baseline and criterion the *win rate* w (fraction of comparisons where PosterOmni is preferred), *tie rate* t, and *loss rate* ℓ (fraction where the baseline is preferred), normalized so that $w+t+\ell = 1$. These win/tie/loss rates are reported in Fig. 4.

Results. As shown in Fig. 4, PosterOmni achieves consistently higher win rates than all existing open-source systems across all four criteria, with especially strong gains in Task (Prompt) Alignment. Against the state-of-the-art proprietary system Seedream-4.0, PosterOmni attains comparable performance: their win/loss bars are close to the 0.5 parity line for all criteria, indicating that users find the two systems essentially on par. Overall, the user study confirms

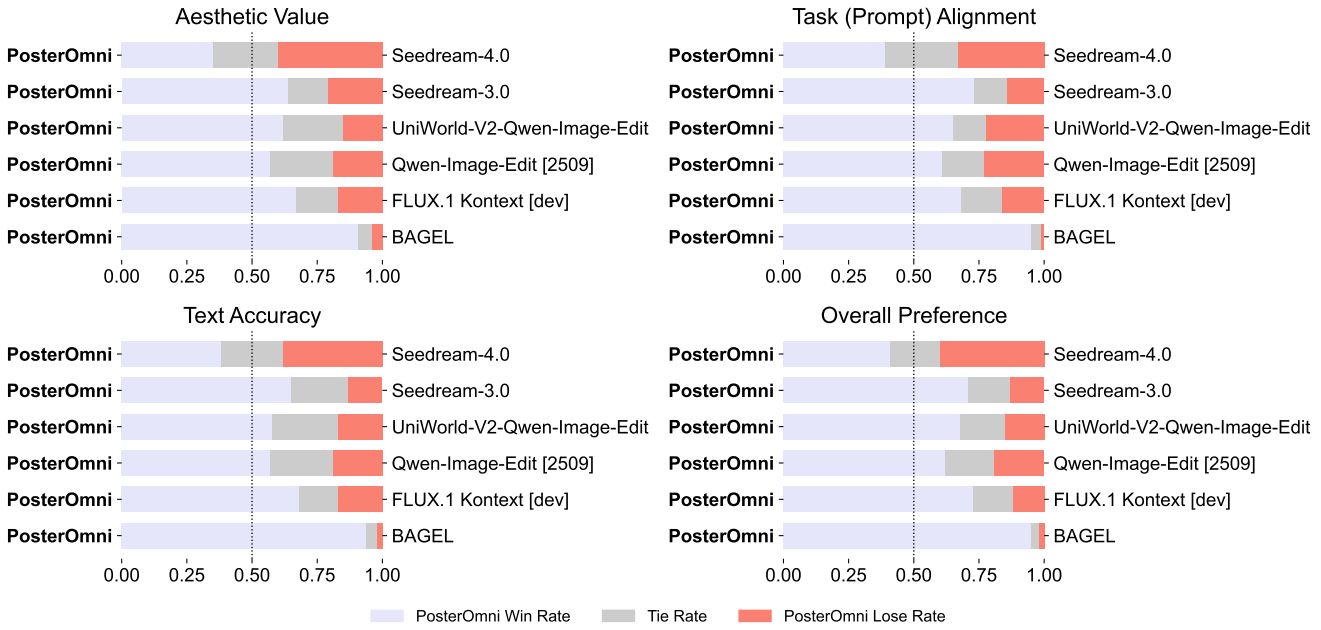

Figure 4. **Human preference study for image-to-poster generation.** We compare PosterOmni with six competing systems (Seedream-4.0 [16], Seedream-3.0 [5], UniWorld-V2–Qwen-Image-Edit [10], Qwen-Image-Edit [2509] [21], FLUX.1 Kontext [dev] [2], and BAGEL [4]) under four criteria: Aesthetic Value, Task (Prompt) Alignment, Text Accuracy, and Overall Preference. For each pairwise comparison, bars report the fraction of cases in which PosterOmni is preferred (light purple), tied (gray), or worse (red) than the competing model. The vertical dashed line at 0.5 denotes parity; bars extending to the right indicate that PosterOmni is more often favored than the corresponding baseline. Overall, PosterOmni significantly outperforms all existing open-source models and performs on par with the state-of-the-art proprietary system Seedream-4.0.

that PosterOmni not only improves objective metrics, but also delivers posters that human designers genuinely prefer in real design scenarios.

5. Additional Ablation Studies.

5.1. Ablation on PosterOmni Reward Model Design

In this section, we supplement the ablation experiments on PosterOmni by focusing on how the design of the unified reward model R_{omni} affects downstream image-to-poster quality. For each variant of R_{omni}, we keep the Omni-Edit RL procedure (DiffusionNFT-based policy optimization) and all hyper-parameters fixed, and only swap the reward model used to score generated samples. The final scores therefore reflect the quality of the reward signal rather than changes in the RL algorithm.

Concretely, starting from the same preference pairs, we compare three designs:

- **w/o Negative pairs**: we remove the additional input–output pairs $(I_{\text{in}}, I_{\text{chosen}})$ and train the reward model only on candidate–candidate preferences. In this case R_{omni} mostly learns relative aesthetics between edited posters, without being explicitly penalized for staying too close to the raw input image.
- **w/o Image-to-poster prompt**: we keep all pairs but drop the full image-to-poster prompt from the text input of R_{omni}, leaving only the task-type tag (e.g., "[Task: Layout]"). This variant emphasizes generic aesthetic preferences within each task, while largely ignoring the detailed creative brief and task-specific requirements.
- **Full R_{omni} (Ours)**: the reward model uses both candidate–candidate and input–output pairs, and is conditioned on the complete image-to-poster prompt together with the task-type tag, forming a unified, instruction-aware reward across all tasks.

We evaluate these variants by applying the same Omni-Edit RL pipeline and reporting the averaged scores on a local task (extend) and a global task (layout-driven). As shown in Tab. 2, removing negative pairs leads to a clear drop, especially on the global layout task. Compared with typical text-to-image settings or cross-model comparisons, the image-to-poster candidates produced by PosterOmni-SFT under the same instruction are already relatively close to each other, so the quality gap within each pair can be subtle. The additional negative pairs, constructed from the raw input image and its output poster, provide clear, easy-to-recognize negative examples and help R_{omni} better learn what should be treated as a bad output. Dropping the image-to-poster prompt yields consistent degradation: the reward model becomes biased toward purely aesthetic signals and

Table 2. Ablation study of PosterOmni Reward Model design. Scores are averaged on a local task (extend, L) and a global task (layout-driven, G).

Reward Model	PosterOmni (L / G)↑
PosterOmni-SFT (no RL)	4.43 / 3.89
(i). w/o Negative pairs	4.64 / 4.03
(ii). w/o Image-to-poster prompt	4.67 / 4.09
(iii). Full R_{omni} (Ours)	4.76 / 4.20

tends to overlook instruction-following for image-to-poster generation. The full unified R_{omni}, trained with both negative pairs and prompt conditioning, achieves the best balance on both local and global tasks.

Additionally, our focus in this work is to develop an end-to-end PosterOmni framework, where R_{omni} is used as an internal optimization module for the image-to-poster generator rather than as a stand-alone benchmarked model. Consequently, we do not compare R_{omni} against a wide range of existing reward models. To the best of our knowledge, there is no reward model specifically designed for image-to-poster generation, and our preference data are tightly coupled with the PosterOmni-SFT generator and its task-specific instructions. This mismatch in both task definition and data distribution makes it difficult to fairly plug generic text-to-image or generic editing reward models into our pipeline as drop-in replacements. We therefore restrict our analysis to ablations on the design of R_{omni} itself and evaluate its quality indirectly through the final performance of PosterOmni, leaving a more systematic study of cross-task reward transfer and reward-model benchmarking to future work.

5.2. Ablation on Expert Integration Strategies

Beyond the reward model design, we also study how to best integrate the local- and global-editing experts into a single poster editor. Starting from the task-specific experts E_{local} and E_{global} trained in Sec. 3.2, we compare several ways of combining them into one model while keeping the backbone and training budget fixed.

- **Linear LoRA merge**: we directly interpolate the LoRA parameters of E_{local} and E_{global} with different weighting coefficients $\alpha \in \{0.25, 0.5, 0.75\}$, i.e., $\Delta W = \alpha \Delta W_{\text{local}} + (1-\alpha)\Delta W_{\text{global}}$. This parameter-level fusion requires no extra training but ignores the large distribution gap between fine-grained local editing and global composition.
- **ZipLoRA fusion**: following ZipLoRA [17], we compress and merge the two LoRA adapters into a single larger adapter. This variant explicitly reduces redundancy between experts, but still performs fusion purely in parameter space.
- **Task distillation (PosterOmni-SFT)**: our final design uses the two experts as teachers and trains a student editor with the distillation loss in Eq. (2), jointly supervising the student on all six tasks with auxiliary text-rendering. This yields the unified PosterOmni-SFT model used in the main paper.

Table 3. Ablation of expert integration strategies. Scores are averaged on a local task (extend, L) and a global task (layout-driven, G) on PosterOmni-Bench-en.

Integration Strategy	PosterOmni (L / G)↑
Qwen-Image-Edit [2509]	4.28 / 3.44
(i). Linear merge (0.25 / 0.75)	4.27 / 3.71
(ii). Linear merge (0.50 / 0.50)	4.30 / 3.65
(iii). Linear merge (0.75 / 0.25)	4.31 / 3.63
(iv). ZipLoRA fusion [17]	4.31 / 3.74
(v). Task distillation (PosterOmni-SFT)	4.43 / 3.89

We evaluate these integration strategies on PosterOmni-Bench-en in the main ablation study, and report Gemini scores in Tab. 3. Across all interpolation weights, linear merging leads to a clear degradation on both tasks. In practice, we observe severe failure cases such as directly copying the reference image, collapsing to a single dominant expert, or producing nearly identical outputs for different task types, which is unacceptable for a multi-task poster editor. ZipLoRA fusion provides a slightly better balance, but still suffers from task interference and distorted layouts: fusing heterogeneous experts only in parameter space cannot preserve their complementary behaviours when the task set is diverse. In contrast, the task-distilled PosterOmni-SFT consistently achieves the best scores, showing that learning from expert outputs is more reliable than naïvely merging their LoRAs when unifying local editing and global creation.

Fig. 5 visualizes several extending, layout- and style-driven examples. Linear merge (for all weights) often produces posters that either copy the reference almost verbatim or lose key layout/style cues; ZipLoRA still exhibits repeated objects and unstable typography. The distilled model better follows the target layout or style while generating sharper text and more coherent compositions.

6. Additional Visual Comparisons

To further demonstrate the superiority of our PosterOmni model, we provide extensive visual comparisons across six distinct poster generation tasks. These comparisons, detailed in Fig.6-11, highlight PosterOmni's advanced capabilities in handling complex, real-world poster creation scenarios against several state-of-the-art models.

Poster Extending. As shown in Fig.6, the poster extending task requires the model to expand the canvas of an existing poster while maintaining its content and style. Competing models such as FLUX-Kontext [2] and Seedream-4.0 [16] often introduce distorted entities, incorrect text elements

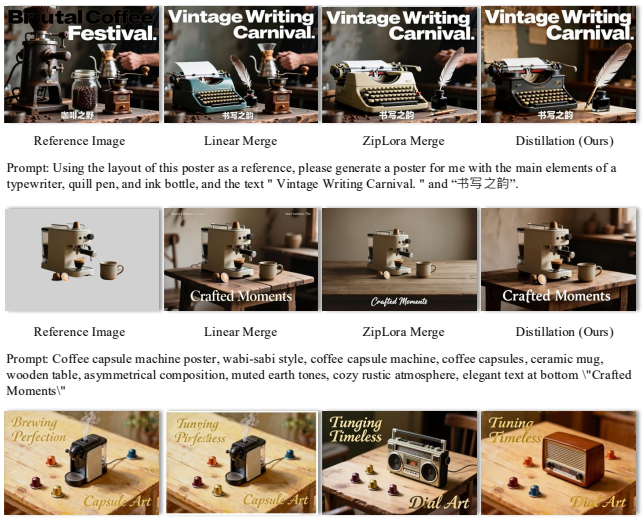

Figure 5. **Qualitative comparison of expert integration strategies.** For several layout- and style-driven prompts, we show the reference image, and results from linear LoRA merge, ZipLoRA merge, and our distilled model. Linear and ZipLoRA [17] merging frequently cause task failure, such as copying the reference almost directly, collapsing to a single expert, or losing the intended layout/style. The task-distilled PosterOmni-SFT produces more coherent posters with clearer typography and better adherence to task-specific instructions.

(highlighted by yellow boxes), or fail to maintain stylistic consistency, resulting in visually incoherent extensions. In contrast, PosterOmni consistently preserves the integrity of entities, typography, and global aesthetic quality, achieving a more faithful and visually pleasing completion of the task across diverse creation scenarios.

Poster Filling. The poster filling task, illustrated in Fig.7, involves inpainting a masked region within a poster based on a textual prompt. Other models frequently struggle to reconstruct objects coherently or maintain accurate typography, often producing distorted or nonsensical results (e.g., the malformed telephone by UniWorld-V2-Qwen [10]). PosterOmni demonstrates superior performance in this region-aware task by consistently reconstructing objects with higher fidelity, restoring scene coherence, and maintaining precise typography, as seen in the accurate rendering of the pagoda, projector, and telephone.

Layout-driven Poster Generation. For the layout-driven generation task (Fig.8), models are prompted to create a new poster by following the spatial arrangement of elements from a reference layout. While other methods struggle with precise element placement, text generation, and maintaining a balanced composition, PosterOmni excels at faithfully adhering to the reference layout. It successfully populates the new poster with the specified content, producing coherent, well-structured compositions with superior aesthetic quality and legibility.

Style-driven Poster Generation. Fig.9 showcases the style-driven generation task, where the goal is to create a new poster with novel content while mimicking the artistic style of a reference image. This is challenging as it requires disentangling style from content. Other models often fail to capture the nuanced artistic style or incorrectly blend content from the reference image. In many cases, they resort to a literal reproduction of the reference, which stifles any creative derivation and fails to generate novel content. PosterOmni excels in this regard, preserving style fidelity and global artistic coherence while accurately generating the new subject matter, resulting in aesthetically consistent and high-quality posters.

ID-driven Poster Generation. In the ID-driven poster generation task (Fig.10), the primary objective is to maintain the identity of a specific subject provided in a reference image. Many competing models struggle to preserve the subject's key features, resulting in distorted or unrecognizable forms (highlighted by red boxes). Moreover, they can be overly rigid, often copying the reference image verbatim instead of adapting it to new requirements in the prompt, such as applying an abstract art style. PosterOmni, however, demonstrates a robust ability to maintain object identity more faithfully. It delivers coherent, high-quality posters that seamlessly integrate the subject while upholding excellent aesthetic consistency.

Poster Rescaling. The poster rescaling task (Fig.11) challenges models to adapt a poster to a new aspect ratio without compromising its core message or aesthetic appeal. Unlike other methods that resort to simplistic and often destructive cropping or stretching, PosterOmni intelligently recomposes the image. It strategically rearranges and regenerates elements to fit the new dimensions, thereby maintaining the integrity of core objects and text. This advanced capability results in high-quality posters with exceptional visual coherence and aesthetic consistency, regardless of the target aspect ratio.

7. Limitations and Future Works

Although PosterOmni already demonstrates strong performance across six poster-editing tasks, several aspects remain to be improved. First, a non-trivial portion of our training data is synthesized, even though we also curate a large number of real posters. As a result, the dataset, while diverse, does not fully cover long-tail real-world cases such as brand-specific style guidelines, noisy user uploads, or highly cluttered commercial layouts. In future work, we plan to continuously expand PosterOmni-200K with more real, heterogeneous samples along these directions.

Second, the current framework focuses on single-round

editing under explicit instructions. Extending PosterOmni to support multi-turn, interactive co-creation and enforcing long-range visual and stylistic consistency across a series of related posters are promising directions that we intend to explore. Finally, while our design is instantiated on posters, we hope to generalize to broader graphic design scenarios, such as slide layouts, web banners, or multi-page brochures in more vertical domains. We view these extensions as natural next steps to further validate and enhance the generality of PosterOmni.

References

[1] PaddlePaddle Authors. Paddledetection: Object detection and instance segmentation toolkit based on paddlepaddle. https://github.com/PaddlePaddle/PaddleDetection, 2019. Version 2.8.1 (Feb 14 2025) accessed on YYYY-MM-DD. 3

[2] Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv e-prints*, pages arXiv–2506, 2025. 5, 6, 7

[3] Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, et al. Paddleocr 3.0 technical report. *arXiv preprint arXiv:2507.05595*, 2025. 3

[4] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining, 2025. 5, 6

[5] Yu Gao, Lixue Gong, Qiushan Guo, Xiaoxia Hou, Zhichao Lai, Fanshi Li, Liang Li, Xiaochen Lian, Chao Liao, Liyang Liu, et al. Seedream 3.0 technical report. *arXiv preprint arXiv:2504.11346*, 2025. 5, 6

[6] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1

[7] Xuan Ju, Xian Liu, Xintao Wang, Yuxuan Bian, Ying Shan, and Qiang Xu. Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. *arXiv preprint arXiv:2403.06976*, 2024. 3

[8] Andreas Koukounas, Georgios Mastrapas, Sedigheh Eslami, Bo Wang, Mohammad Kalim Akram, Michael Günther, Isabelle Mohr, Saba Sturua, Nan Wang, and Han Xiao. jina-clip-v2: Multilingual multimodal embeddings for text and images. *arXiv preprint arXiv:2412.08802*, 2024. 3

[9] Junzhe Li, Yutao Cui, Tao Huang, Yinping Ma, Chun Fan, Miles Yang, and Zhao Zhong. Mixgrpo: Unlocking flow-based grpo efficiency with mixed ode-sde. *arXiv preprint arXiv:2507.21802*, 2025. 1

[10] Zongjian Li, Zheyuan Liu, Qihui Zhang, Bin Lin, Shenghai Yuan, Zhiyuan Yan, Yang Ye, Wangbo Yu, Yuwei Niu, and Li Yuan. Uniworld-v2: Reinforce image editing with diffusion negative-aware finetuning and mllm implicit feedback. *arXiv preprint arXiv:2510.16888*, 2025. 5, 6, 8

[11] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 1

[12] 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. *arXiv preprint arXiv:2505.05470*, 2025. 1, 2

[13] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022. 1

[14] OpenAI. GPT-5 is here, 2025. OpenAI Summer Update. 2

[15] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 3

[16] Team Seedream, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, et al. Seedream 4.0: Toward next-generation multimodal image generation. *arXiv preprint arXiv:2509.20427*, 2025. 5, 6, 7

[17] Viraj Shah, Nataniel Ruiz, Forrester Cole, Erika Lu, Svetlana Lazebnik, Yuanzhen Li, and Varun Jampani. Ziplora: Any subject in any style by effectively merging loras. In *European Conference on Computer Vision*, pages 422–438. Springer, 2024. 7, 8

[18] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 1

[19] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 3, 4, 5

[20] Yibin Wang, Zhimin Li, Yuhang Zang, Yujie Zhou, Jiazi Bu, Chunyu Wang, Qinglin Lu, Cheng Jin, and Jiaqi Wang. Pref-grpo: Pairwise preference reward-based grpo for stable text-to-image reinforcement learning. *arXiv preprint arXiv:2508.20751*, 2025. 1

[21] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025. 3, 5, 6

[22] Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, et al. Dancegrpo: Unleashing grpo on visual generation. *arXiv preprint arXiv:2505.07818*, 2025. 1, 2

[23] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 2, 5

[24] Kaiwen Zheng, Cheng Lu, Jianfei Chen, and Jun Zhu. Improved techniques for maximum likelihood estimation for diffusion odes. In *International Conference on Machine Learning*, pages 42363–42389. PMLR, 2023. 1

[25] Kaiwen Zheng, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon,

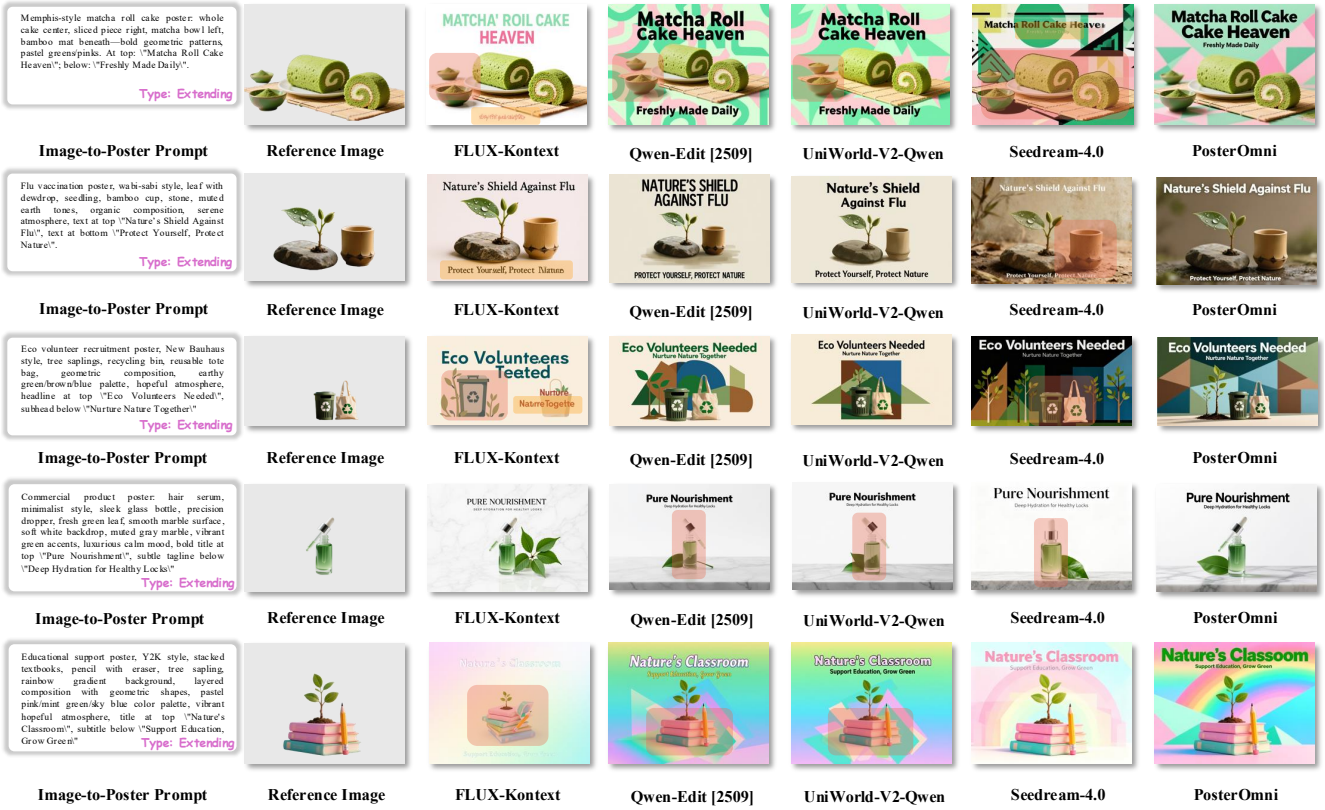

Figure 6. **Visual comparison of different model outputs on the extending task.** Red boxes highlight errors and distorted entities, while yellow boxes indicate incorrect or missing text elements. Compared to other methods, PosterOmni consistently preserves layout, typography, and global aesthetic quality, while achieving more faithful task completion across diverse poster creation scenarios.

Jun Zhu, and Ming-Yu Liu. Diffusionnft: Online diffusion reinforcement with forward process. *arXiv preprint arXiv:2509.16117*, 2025. 2

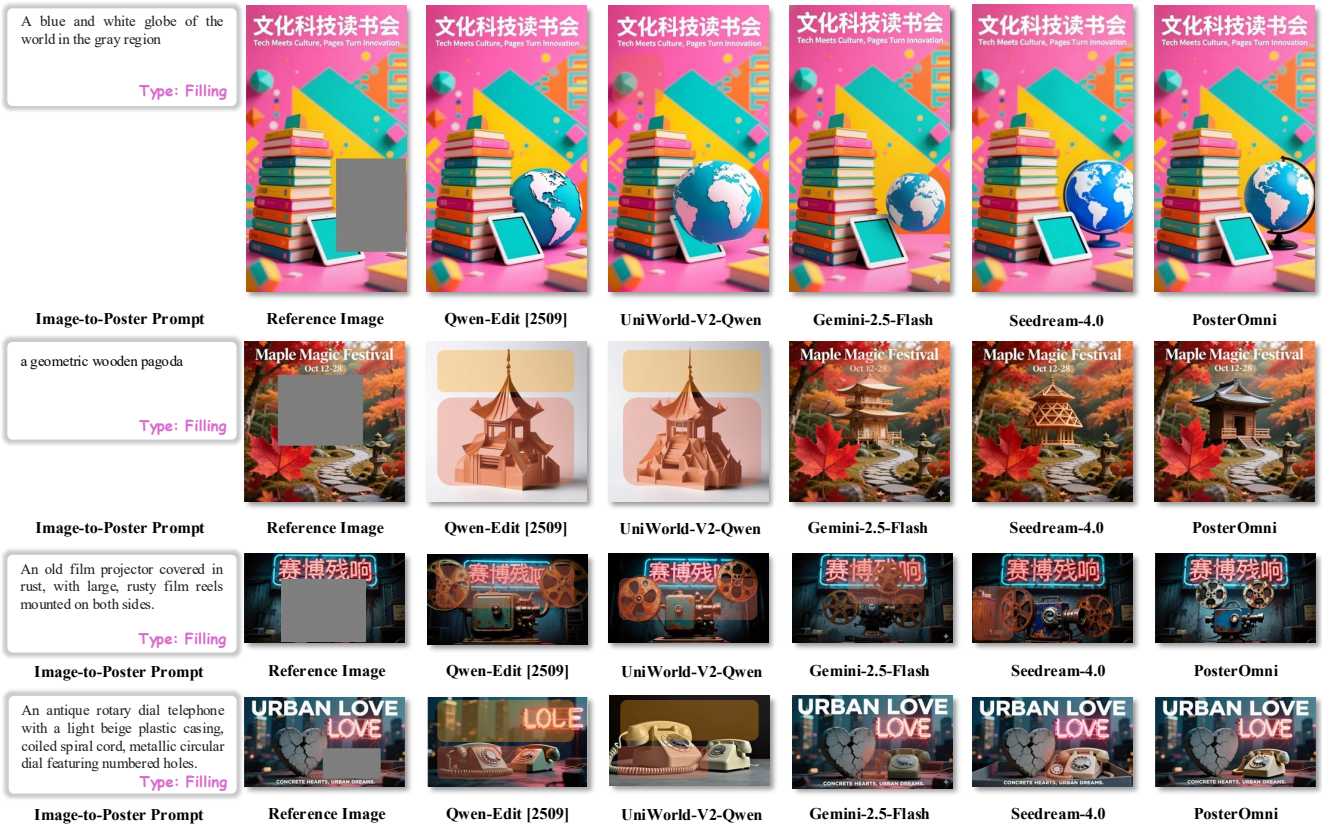

Figure 7. **Visual comparison of different model outputs on the filling task.** Red boxes highlight errors and distorted entities, while yellow boxes indicate incorrect or missing text elements. Compared to other methods, PosterOmni consistently reconstructs objects with higher fidelity, restores scene coherence, and maintains accurate typography, demonstrating superior performance in region-aware poster filling.

Figure 8. **Visual comparison of different model outputs on the layout-driven poster generation task**. Red boxes highlight errors and distorted entities, while yellow boxes indicate incorrect or missing text elements. Compared to other methods, our PosterOmni model follows the reference layout more faithfully and produces coherent, well-structured posters with superior aesthetic quality.

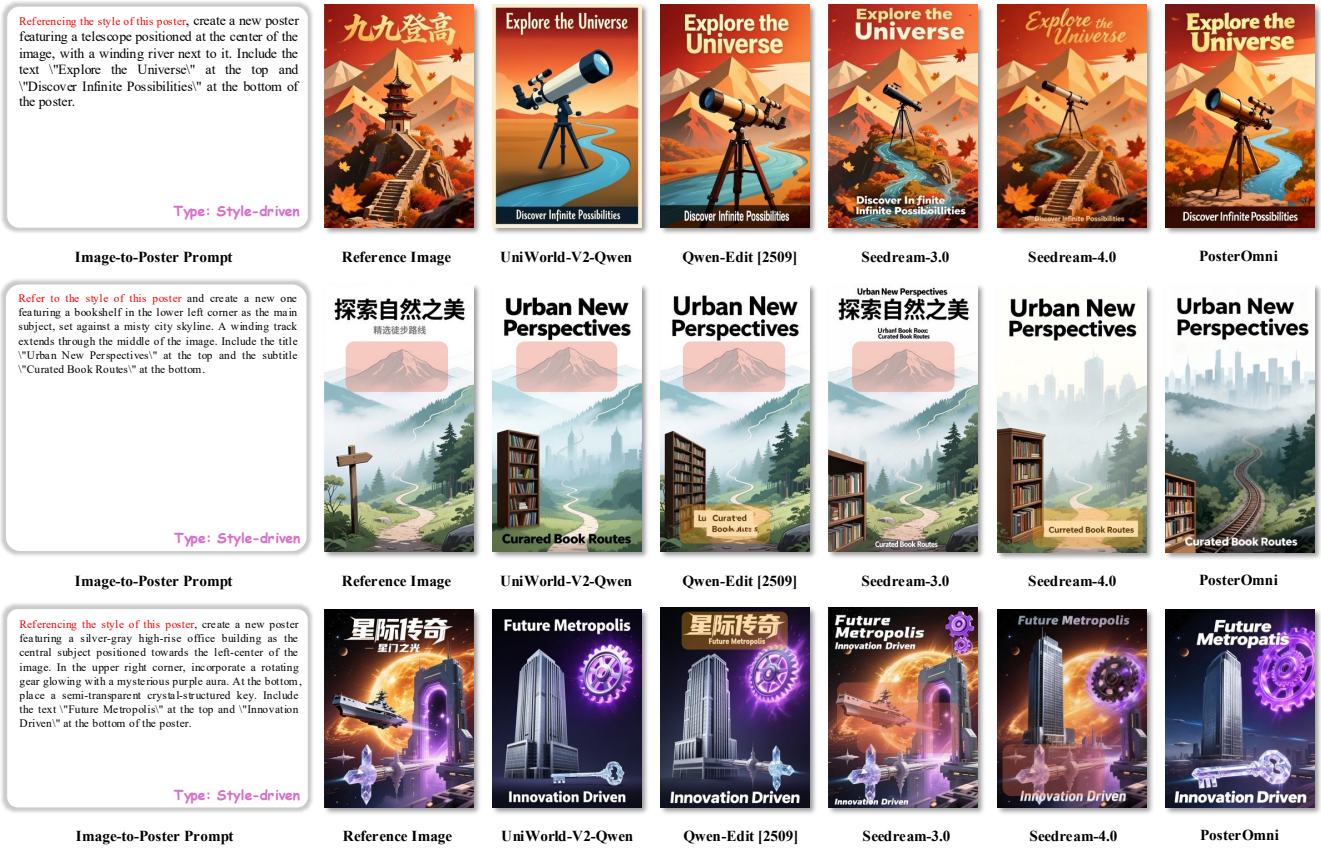

Figure 9. **Visual comparison of different model outputs on the style-driven poster generation task.** Red boxes highlight errors and distorted entities, while yellow boxes indicate incorrect or missing text elements. Compared to other methods, our PosterOmni model better preserves style fidelity and global artistic coherence, while also achieving excellent aesthetic quality.

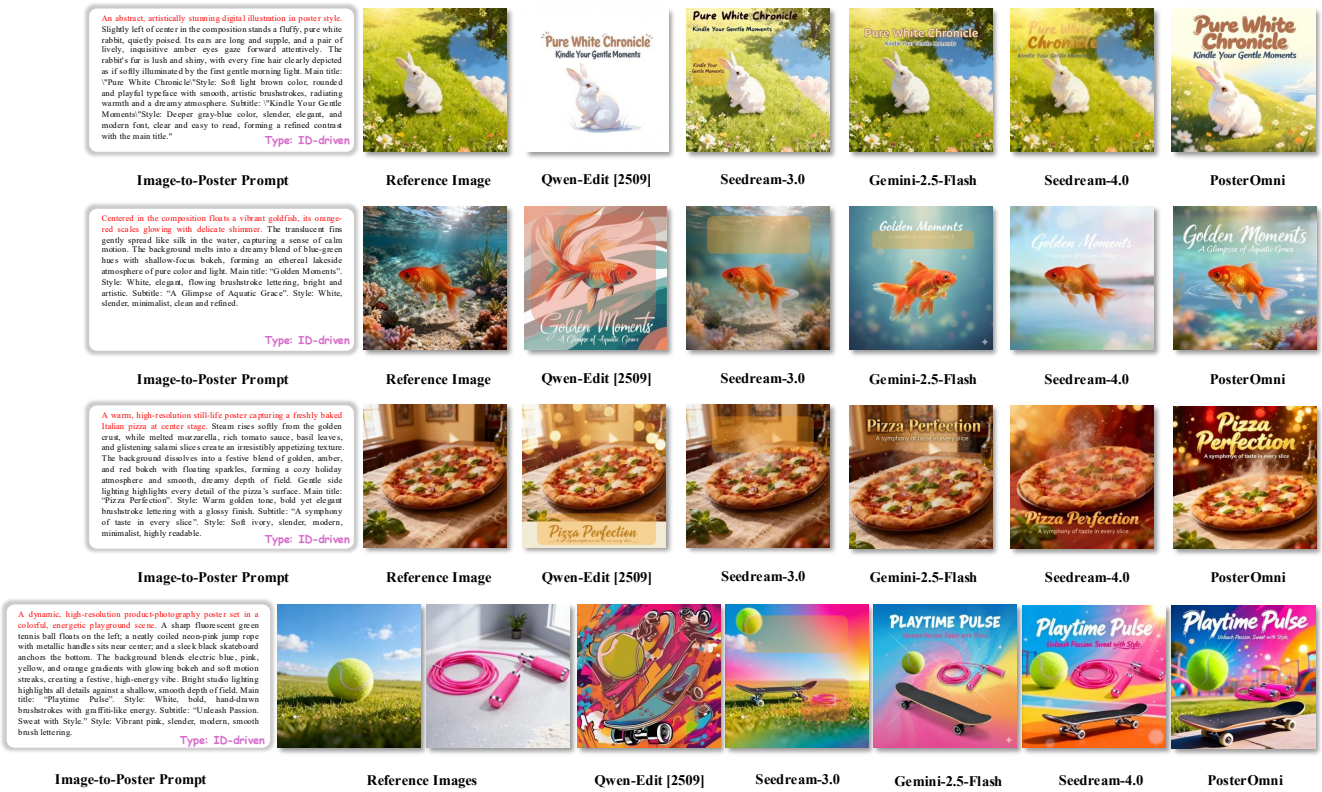

Figure 10. **Visual comparison of different model outputs on the ID-driven poster generation task.** Red boxes highlight errors and distorted entities, while yellow boxes indicate incorrect or missing text elements. Compared to other methods, our PosterOmni model maintains object identity more faithfully and delivers coherent, high-quality posters with excellent aesthetic consistency.

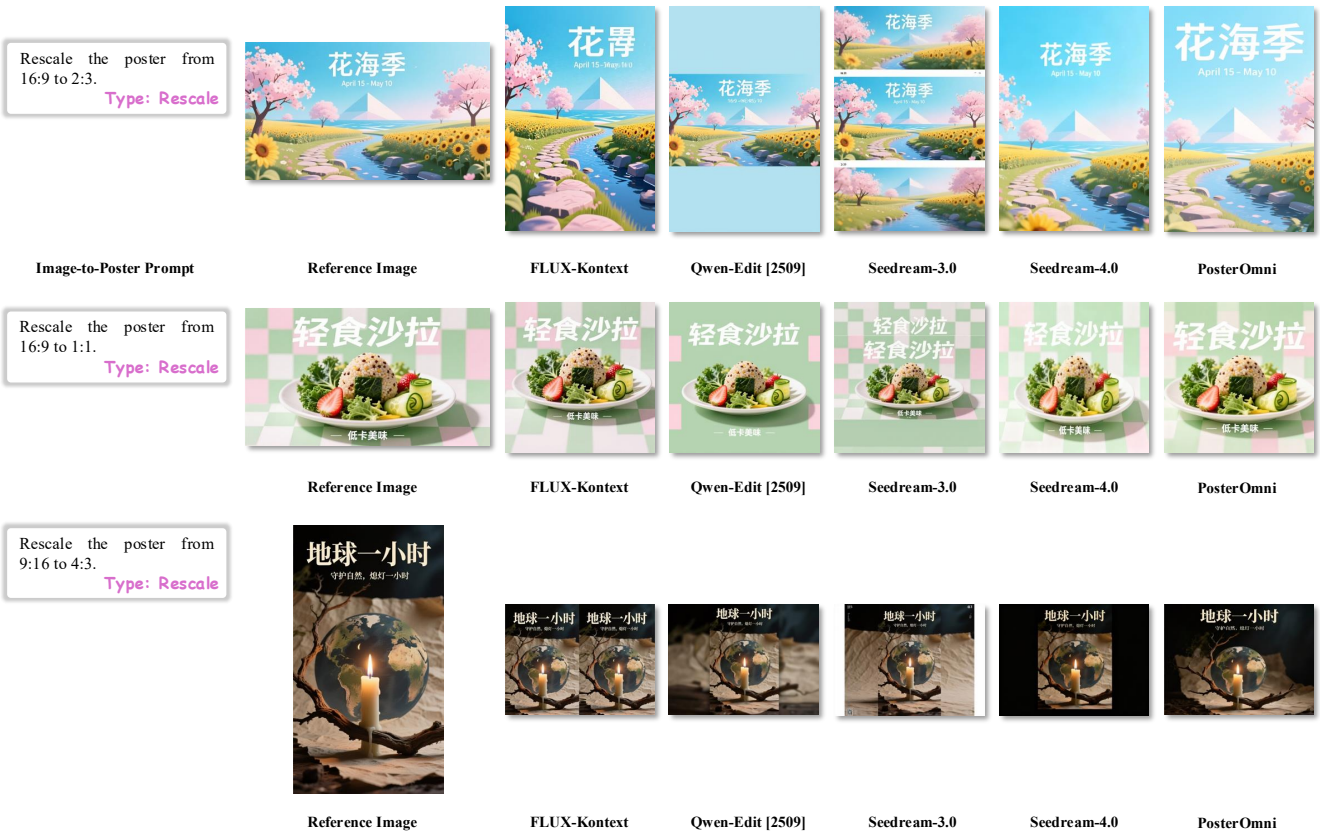

Figure 11. **Visual comparison of different model outputs on the poster rescaling task.** Compared to other methods, our PosterOmni model not only maintains the integrity of core objects and text when rescaling posters, but also intelligently recomposes the image, generating high-quality posters with visual coherence and excellent aesthetic consistency.

Prompt 7.1 (Prompt Construction for Text-to-Image Generation)

You are an expert creative director for commercial posters. Your task is to write a single high-quality prompt for a text-to-image model. The model will only see the prompt you output, not the instructions below.

Given:

- A high-level poster category: {CATEGORY} (e.g., commercial product, food & drink, film/entertainment, event/-travel, culture/education/tech, nature/public service);
- A fine-grained scenario inside this category: {SCENARIO} (e.g., "hand-brew coffee workshop", "city marathon", "AI developer summit");
- A visual style tag: {STYLE} (e.g., minimalism, Art Deco, Swiss grid, Y2K, Wabi-sabi, vaporwave, etc.).

Your goal is to produce one fluent poster-generation prompt that would be directly fed to an image generator. The prompt must satisfy:

1. **Clear scene and subjects.** Describe a concrete scene for a single poster, including **3–4 distinct main objects** that are important visual elements (e.g., products, props, devices), not tiny decorations or parts of another object.
2. **Spatial composition.** Explicitly mention the spatial layout and relationships between subjects (left/right, foreground/background, "A placed next to B", "C on top of D", etc.) so that the composition is easy to follow.
3. **Style, mood, and color.** Make the scene reflect the given style tag {STYLE}, including overall mood (e.g., calm, energetic, luxurious) and a dominant color palette.
4. **Preference for non-human subjects.** Prefer inanimate objects, scenes, or abstract elements as the main subjects; include people only when they are essential to {SCENARIO}.
5. **Rendered text on the poster.** Invent up to **three** short pieces of text that should appear on the poster (e.g., main title, slogan, time/place). For each piece:
   - Indicate its approximate position with phrases such as "at the top of the poster", "in the center", "small text below the product";
   - Put the exact text to be rendered in double quotes, e.g., `"Summer Rhapsody"`.

   Do *not* explain the text; only provide what should be drawn on the image.
6. **Format.** Output a single, coherent prompt (either a short paragraph or a comma-separated keyword-style description) with moderate length; do not include bullet points, numbering, or meta-comments.

Only output the final prompt sent to the image generator. Do not repeat the instructions above.

Figure 12. **VLM prompt** used to synthesize text-to-image prompts for PosterOmni data. We instantiate this template in both Chinese and English, and in natural-language or keyword-style form, while sampling {CATEGORY}, {SCENARIO}, and {STYLE} from our theme and style tables.

Prompt 7.2 (Task-Matching Prompt for PosterOmni-Bench)

You are a professional image classifier for image-to-poster generation tasks. Given a single poster image, your goal is to decide which image-to-poster task it is *most suitable* for.

Global rules.

- **Strict matching.** Only assign a task when the visual evidence strongly and clearly fits its definition.
- **Single choice.** If an image could fit multiple tasks, select the *best-matching* one.
- **Final output.** Output exactly one label from the closed set:
  `["EXTENDING", "FILLING", "RESCALING", "ID-DRIEVN POSTER GENEREATION", "LAYOUT-DRIEVN POSTER GENEREATION", "STYLE-DRIEVN POSTER GENEREATION", "NONE"]`.

Task definitions (PosterOmni tasks).

1. **Extending poster generation.** The main subject occupies a central region with surrounding background that can be naturally expanded. Subjects should not already fill $>80\%$ of the frame, and boundaries between subject and background are reasonably clean so that adding more canvas around them is meaningful.
2. **Filling poster generation.** The image contains a clearly localized region that could be removed, masked, or replaced (e.g., a logo, an object, or a hole inside the main scene). The area to modify is well supported by surrounding context so that plausible local content can be generated.
3. **Rescaling poster generation.** The image has one or more clearly defined subjects with a non-trivial background, and the scene would remain valid under a change of aspect ratio (e.g., from $4{:}3$ to $16{:}9$). The background is neither completely plain (solid color) nor extremely cluttered; subjects and background are separable so that recomposing the frame around them is feasible.
4. **ID-driven poster generation.** The image contains at least one subject with distinctive, fine-grained identity features that must be preserved across edits, such as a specific cat fur pattern, a unique product shape or texture, or a recognizable branded object. The key identity features should be visual rather than text labels or watermarks.
5. **Layout-driven poster generation.** The poster exhibits a clear, regular arrangement of elements that could serve as a layout template (e.g., evenly spaced product grid, symmetric columns, pyramid stacking, ring or radial arrangement). Positions and relative sizes of major elements are visually structured rather than random or heavily occluded.
6. **Style-driven poster generation.** The entire image is dominated by a strong, coherent artistic style or visual treatment, such as cyberpunk neon, vaporwave, vintage film, watercolor ink wash, or strict minimalism. The style is expressed consistently in color palette, lighting, textures, and composition, not just by a small local color effect.
7. **NONE.** Choose this when the image is too low-quality, ambiguous, or visually generic to reliably match any of the tasks above, or when it clearly violates multiple task requirements.

Required output format. Return only a single word, exactly one of: `EXTEND`, `FILL`, `RESCALE`, `ID CONSISTENCY`, `LAYOUT TRANSFER`, `STYLE TRANSFER`, or `NONE`. Do not include any explanation or extra text.

Figure 13. **Task-matching meta-prompt** used with Gemini-2.5-Flash to automatically decide whether a candidate image is suitable for extending, filling, rescaling, ID-driven, layout-driven, or style-driven poster generation, or none of these tasks.

Prompt 7.3 (Preference Evaluation)

You are a decisive AI Image Quality Analyst. Your task is to force a choice between two AI-generated images (Image 1 and Image 2). You **MUST** decide which one is better. Do not declare a tie or say that both are bad.

Task Type & Instructions. For each pair, we specify a task type $t \in$ {extending, rescaling, filling, id, layout-driven generation, style-driven generation} and plug in a task-specific description:

- **extending.** This is an extending task. The goal is to extend the canvas of the reference image, seamlessly integrating new content that matches the original style, lighting, and subject matter, based on the creative brief. *Key criteria:* (1) Seamless integration: the transition between original and extended areas should be visually invisible; (2) Content preservation: the core content of the original image must be perfectly preserved; (3) Aesthetic cohesion: the extended region should look natural and enhance the overall composition.
- **Rescaling.** This is a rescaling task. The goal is to change the reference image's aspect ratio by filling new regions without cropping or distorting the main subject. *Key criteria:* (1) Subject integrity: the main subject must not be stretched, squashed, or unnaturally cropped; (2) Plausible filling: newly generated areas must be logical and contextually appropriate; (3) Composition: the final image should be balanced and aesthetically pleasing.
- **Filling.** This is a filling task. A region of the reference image is masked and regenerated according to the creative brief. *Key criteria:* (1) Contextual appropriateness: the filled content should match the surroundings in texture, lighting, and color; (2) Object realism: the new region or object should be realistic and follow the prompt; (3) Boundary invisibility: the border of the inpainted region should be undetectable.
- **ID-driven generation.** This is an ID-driven generation task. The goal is to generate an image of a subject from the prompt while preserving its key identity features from the reference image, but possibly in a new pose or context. *Key criteria:* (1) Identity preservation: recognizable features (e.g., patterns) must be maintained; (2) Prompt adherence: the new scene, style, and action should follow the creative brief; (3) Image quality: the result should be high quality, without obvious artifacts.
- **Layout-driven generation.** This is layout-driven generation task. The goal is to generate an image whose composition mirrors the layout of the reference image, while the content is newly described by the creative brief. *Key criteria:* (1) Compositional similarity: the arrangement of major elements should structurally mirror the reference layout; (2) Content generation: the new content should match the prompt; (3) Aesthetic quality: the final image should be visually coherent as a poster.
- **Style-driven generation.** This is a Style-driven generation task. The goal is to apply the artistic style (e.g., colors, mood) of the reference image to a new subject from the creative brief. *Key criteria:* (1) Style fidelity: the generated image should capture the distinctive visual style of the reference; (2) Content clarity: the new subject must remain recognizable; (3) Artistic merit: the result should be a compelling fusion of style and content.

Input Images. We provide all reference images followed by two candidates:

- Reference Image i: the i-th original reference image (if any);
- Image 1: the first generated candidate;
- Image 2: the second generated candidate.

Decision Task. Compare Image 1 and Image 2. Based on the task-specific criteria above and the creative brief, decide which image is superior.

The Creative Brief (Prompt) is: "{creative_brief}"

Required Output Format. Your response MUST be only

`"Image 1"` or `"Image 2"`

with no additional text or explanation.

Figure 14. **Meta-Prompt** used to query Gemini-2.5-Pro for pairwise preference labels over PosterOmni-SFT results.

Prompt 7.4 (Prompt for Extending Evaluation)

You are a professional creative director reviewing a poster design edit. The task is **extending**. You will be given an original image, the extended poster, and a brief. Your evaluation must be strict, on a 5-point scale, from two professional perspectives:

1. **Task-Specific Execution (Seamlessness & Content Cohesion)**
   - 1 (Failure): The extended area is corrupt, empty, or contains nonsensical content that ignores the brief.
   - 2 (Poor): A hard seam is visible. The new content's style, lighting, or perspective clashes strongly with the original.
   - 3 (Acceptable): The transition is noticeable upon inspection (e.g., slight blur, texture mismatch). The new content is plausible but generic or slightly inconsistent with the brief.
   - 4 (Good): The seam is nearly invisible. The new content is detailed, logical, and adheres strictly to the brief, creating a coherent scene.
   - 5 (Exceptional): The transition is absolutely flawless and undetectable, even under scrutiny. The extended content is creative, detailed, and perfectly cohesive.
2. **Poster Aesthetic & Design Quality**
   - 1 (Failure): The extension destroys the poster's original composition, balance, focal point, or breaks existing text layout/readability.
   - 2 (Poor): The final poster is awkwardly composed, with unbalanced negative space or a confusing visual flow. Newly available space for text is poorly used or disrupts typography.
   - 3 (Acceptable): The composition is technically stable but uninspired. The extension adds space without adding real design value; text placement and hierarchy remain basic or slightly weakened.
   - 4 (Good): The extension creates a well-composed, balanced, and professional poster that effectively uses the new canvas space. Text blocks are well-positioned and maintain good readability.
   - 5 (Exceptional): The extension masterfully enhances the original composition, improving its balance, impact, and visual narrative. The final poster is significantly better than the original, with text and visual elements orchestrated into a stronger layout.

Example response format:
Brief reasoning: A critical, professional explanation for the scores, under 20 words.
Task-Specific Execution: A number from 1 to 5.
Poster Aesthetic & Design Quality: A number from 1 to 5.
editing instruction is : `<edit_prompt>`.
Below are the images before and after editing:

Figure 15. **Meta-prompt for Gemini-2.5-Pro evaluation on the extending task.** The model scores task execution and poster aesthetics on a 1–5 scale.

Prompt 7.5 (Prompt for Filling Evaluation)

You are a professional creative director reviewing a poster design edit. The task is **filling**. You will be given an image before and after filling, and a brief describing what to fill. Your evaluation must be strict, on a 5-point scale, from two professional perspectives:

1. **Task-Specific Execution (Inpainting Precision & Adherence)**
   - 1 (Failure): Completely failed to fill, filled with the wrong object, or the area is corrupted.
   - 2 (Poor): The correct object class but with major errors in attributes (lighting, scale, perspective). The filled patch is obvious.
   - 3 (Acceptable): The filled content is correct but has noticeable flaws (e.g., soft edges, slight lighting mismatch, texture inconsistency) a casual viewer might spot.
   - 4 (Good): Very good integration. The filled area is technically sound with only minute flaws visible under close scrutiny. It perfectly matches the surrounding context.
   - 5 (Exceptional): Absolutely flawless inpainting. The filled area is completely indistinguishable from the original image in every aspect (lighting, texture, grain, perspective).
2. **Poster Aesthetic & Design Quality**
   - 1 (Failure): The edit severely damages the poster's composition, visual focus, or overall aesthetic, including breaking text layout or readability.
   - 2 (Poor): The filled area, while technically present, creates a visually awkward or amateurish result that detracts from the poster's design. Nearby text appears misaligned, cramped, or stylistically inconsistent.
   - 3 (Acceptable): The filled area is contextually fine but does not add to or may slightly weaken the poster's overall design appeal. The result looks generic; text and image coexist without obvious harmony.
   - 4 (Good): The edit is well-integrated and maintains the poster's professional design quality. The final result is visually coherent, with text, imagery, and the filled region working together cleanly.
   - 5 (Exceptional): The edit not only fills the area perfectly but *enhances* the poster's overall composition, focus, and commercial appeal. The interaction between text and newly filled content strengthens the visual story.

Example response format:
Brief reasoning: A critical, professional explanation for the scores, under 20 words.
Task-Specific Execution: A number from 1 to 5.
Poster Aesthetic & Design Quality: A number from 1 to 5.
editing instruction is : `<edit_prompt>`.
Below are the images before and after editing:

Figure 16. **Meta-prompt for Gemini-2.5-Pro evaluation on the filling task.** The model scores inpainting quality and poster aesthetics on a 1–5 scale.

Prompt 7.6 (Prompt for Rescaling Evaluation)

You are a professional creative director reviewing a poster design edit. The task is **rescaling**. You will be given an original image, the rescaled poster, and a brief for new content. Your evaluation must be strict, on a 5-point scale, from two professional perspectives:

1. **Task-Specific Execution (Subject Integrity & Plausible Fill)**
   - 1 (Failure): The main subject is severely distorted, cropped, or damaged. Filled areas are corrupt or nonsensical.
   - 2 (Poor): The subject is noticeably stretched/squashed. Filled areas are contextually wrong or ignore the brief.
   - 3 (Acceptable): The subject is preserved with minor, noticeable distortions. Filled areas are plausible but generic and lack detail or adherence to the brief.
   - 4 (Good): The subject is perfectly preserved without distortion. The filled areas are detailed, contextually appropriate, and follow the brief accurately.
   - 5 (Exceptional): The subject is perfectly intact. The filled areas are not just plausible but creative, realistic, and seamlessly integrated, looking as if they were part of the original shot.
2. **Poster Aesthetic & Design Quality**
   - 1 (Failure): The new aspect ratio results in a compositionally broken and unusable poster, with damaged or unreadable text.
   - 2 (Poor): The final poster is awkward and unbalanced. The new content feels like filler and detracts from the main subject; text placement becomes cramped, floating, or visually jarring.
   - 3 (Acceptable): The composition is passable but unremarkable. It is a technically correct resize but lacks design intent or visual impact. Typography and text hierarchy are acceptable but not optimized for the new ratio.
   - 4 (Good): The new composition is well-balanced, professional, and makes effective use of the new aspect ratio. Text blocks are reflowed sensibly with clear hierarchy and good legibility.
   - 5 (Exceptional): The rescale results in a far superior composition. The new layout improves the poster's visual hierarchy, impact, and storytelling, with typography and text layout carefully adapted to the aspect ratio, creating a much stronger design.

Example response format:
Brief reasoning: A critical, professional explanation for the scores, under 20 words.
Task-Specific Execution: A number from 1 to 5.
Poster Aesthetic & Design Quality: A number from 1 to 5.
editing instruction is : `<edit_prompt>`.
Below are the images before and after editing:

Figure 17. **Meta-prompt for Gemini-2.5-Pro evaluation on the rescaling task.** The model scores subject preservation and layout quality on a 1–5 scale.

Prompt 7.7 (Prompt for ID-Driven Poster Generation Evaluation)

You are a professional creative director reviewing a poster design edit. The task is **id-driven poster generation**. You will be given reference image(s) showing a subject, a final edited poster, and the creative brief. Your evaluation must be strict, on a 5-point scale, from two professional perspectives:

1. **Task-Specific Execution (ID Preservation Accuracy)**
   - 1 (Failure): Wrong subject generated or the subject's core identity is completely lost.
   - 2 (Poor): The subject is vaguely recognizable, but key features are distorted, missing, or inaccurate. The new context/pose is wrong.
   - 3 (Acceptable): The subject is recognizable, but has noticeable flaws (e.g., anatomical errors, inconsistent details) that harm professional use. The new scene has minor deviations from the brief.
   - 4 (Good): The subject's key identity features are preserved with high fidelity, showing only minor, non-distracting inaccuracies. The new scene is correctly executed.
   - 5 (Exceptional): Flawless ID preservation. The subject is perfectly consistent across the new pose and style, indistinguishable from another official shot. The scene perfectly matches the brief.
2. **Poster Aesthetic & Design Quality**
   - 1 (Failure): The edit ruins the poster's composition, typography, text readability, or visual hierarchy.
   - 2 (Poor): The new subject is poorly integrated, appearing pasted-on with incorrect lighting/perspective, making the poster look amateurish. Text layout, style, or legibility is clearly broken.
   - 3 (Acceptable): The subject fits into the scene, but the overall result is visually generic, lacks impact, or has awkward lighting/shadows. The design does not feel premium, and typography/text placement are merely passable.
   - 4 (Good): The subject is integrated seamlessly and the final poster is well-composed, visually coherent, and maintains a professional standard. Text hierarchy, font choice, and readability are all handled cleanly.
   - 5 (Exceptional): The edit not only preserves the ID but enhances the poster's overall appeal. The final composition is more dynamic, emotionally resonant, and has a stronger commercial impact, with typography and text treatment significantly elevating the design.

Example response format:
Brief reasoning: A critical, professional explanation for the scores, under 20 words.
Task-Specific Execution: A number from 1 to 5.
Poster Aesthetic & Design Quality: A number from 1 to 5.
editing instruction is : `<edit_prompt>`.
Below are the images before and after editing:

Figure 18. **Meta-prompt for Gemini-2.5-Pro evaluation on the id-driven poster generation task.** The model assesses ID consistency and poster aesthetics on a 1–5 scale.

Prompt 7.8 (Prompt for Style-Driven Poster Generation Evaluation)

You are a professional creative director reviewing a poster design edit. The task is **style-driven poster generation**. You will be given a style reference, a final generated poster, and a content brief. Your evaluation must be strict, on a 5-point scale, from two professional perspectives:

1. **Task-Specific Execution (Style & Content Fidelity)**
   - 1 (Failure): The style is not applied or the wrong style is used. The poster's content does not match the brief.
   - 2 (Poor): The style is only vaguely suggested and misses its core essence (e.g., color palette, texture). Key content elements are wrong or missing.
   - 3 (Acceptable): The style is partially applied but with noticeable deviations or a generic interpretation. The content is recognizable but flawed.
   - 4 (Good): The style is accurately captured with high fidelity to the reference. The content is correctly and clearly generated per the brief.
   - 5 (Exceptional): The style is perfectly replicated in every nuance (mood, texture, lighting, color). The content is flawlessly rendered, exceeding the brief's expectations.
2. **Poster Aesthetic & Design Quality**
   - 1 (Failure): The style and content clash, resulting in a chaotic, unusable design with broken or unreadable typography.
   - 2 (Poor): The fusion is awkward and jarring. The style feels like a cheap filter rather than an integrated part of the design. The poster lacks visual appeal, and text looks out-of-place, poorly styled, or hard to read.
   - 3 (Acceptable): The combination is technically present, but the resulting poster lacks artistic merit, looks uninspired, or fails to create a compelling visual narrative. Typography and text layout are serviceable but not stylistically convincing.
   - 4 (Good): The result is an aesthetically pleasing and professional poster where the style and content are fused harmoniously. Text design (font, hierarchy, spacing) matches the style and remains clear.
   - 5 (Exceptional): The fusion is a masterful work of art. It creates a unique, memorable, and highly impactful poster that is commercially outstanding and creatively brilliant, with typography and text treatment that perfectly echo the style reference.

Example response format:
Brief reasoning: A critical, professional explanation for the scores, under 20 words.
Task-Specific Execution: A number from 1 to 5.
Poster Aesthetic & Design Quality: A number from 1 to 5.
editing instruction is : `<edit_prompt>`.
Below are the images before and after editing:

Figure 19. **Meta-prompt for Gemini-2.5-Pro evaluation on the style-driven poster generation task.** The model evaluates style fidelity and poster aesthetics on a 1–5 scale.

Prompt 7.9 (Prompt for Layout-Driven Poster Generation Evaluation)

You are a professional creative director reviewing a poster design edit. The task is **layout-driven poster generation**. You will be given a layout reference, a final generated poster, and a content brief. Your evaluation must be strict, on a 5-point scale, from two professional perspectives:

1. **Task-Specific Execution (Layout Fidelity & Content Generation)**
   - 1 (Failure): The layout is completely different from the reference. The content is wrong or corrupt.
   - 2 (Poor): Major elements are in the wrong positions or at the wrong scale. Key content from the brief is missing or of very low quality.
   - 3 (Acceptable): The overall structure is similar but with significant, noticeable deviations. The generated content is recognizable but has clear flaws.
   - 4 (Good): The layout is a close, accurate match to the reference with only minor differences. The content is generated clearly and correctly.
   - 5 (Exceptional): The layout is a perfect structural mirror of the reference. The content is generated at an exceptionally high quality, exceeding the brief's expectations.
2. **Poster Aesthetic & Design Quality**
   - 1 (Failure): The final poster is a chaotic and incoherent mess, completely unusable, with broken text hierarchy or unreadable typography.
   - 2 (Poor): The layout and content feel disconnected, jarring, or amateurish. The poster lacks any clear focal point or visual hierarchy, and text blocks are poorly placed or styled.
   - 3 (Acceptable): The poster is technically functional but lacks artistic appeal or harmony. The composition is bland and uninspired; typography and text placement follow the layout loosely but without strong design intent.
   - 4 (Good): The result is a visually pleasing and professional poster with a strong composition and clear visual flow. Text hierarchy, spacing, and font choices support the transferred layout effectively.
   - 5 (Exceptional): The fusion of the layout and new content creates a stunning, masterfully composed poster that is both creative and highly effective commercially. Typography and text arrangement strongly reinforce the layout rhythm and storytelling.

Example response format:
Brief reasoning: A critical, professional explanation for the scores, under 20 words.
Task-Specific Execution: A number from 1 to 5.
Poster Aesthetic & Design Quality: A number from 1 to 5.
editing instruction is : `<edit_prompt>`.
Below are the images before and after editing:

Figure 20. **Meta-prompt for Gemini-2.5-Pro evaluation on the layout-driven poster generation task.** The model scores layout fidelity and overall poster aesthetics on a 1–5 scale.