E-comIQ-ZH: A Human-Aligned Dataset and Benchmark for Fine-Grained Evaluation of E-commerce Posters with Chain-of-Thought

Supplementary Material

A. Dataset: E-comIQ-18k Details

A.1. Source Composition and Splits

E-comIQ-18k contains 18k images drawn from six sources (See Figure 4 in Sec. 3.1). The proportions are 27.8% merchant HQ, 27.8% merchant LQ, 16.7% open-source posters, 11.1% AI-generated posters, 11.1% AI-edited posters, and 5.6% professional designs.

Merchant originals (HQ / LQ). We start from a large pool of merchant-provided product photos collected from real online listings. Each image is labelled by experts with a binary High Quality(HQ) / Low Quality(LQ) according to overall commercial usability, including product visibility, background cleanliness, text legibility, and layout. Then we randomly sample 5k HQ and 5k LQ images, removing obvious near-duplicates. This procedure gives a broad and realistic quality spectrum for in-the-wild merchant content.

Open-source posters. To increase diversity in style and category coverage, we further sample 3k posters from a public e-commerce poster dataset released in Autoposter [4]. These images are usually complete posters with designed

AI-generated posters. The AI-generated subset is created from product cutouts on a white background. For each product we construct a text prompt that specifies the scene, style, and key selling points, then use GPT-4o as a text-to-image generator conditioned on the cutout as visual reference. Generation prompts and examples are provided in Fig 3, and we discard obvious failures such as missing products or unreadable text.

AI-edited posters. The AI-edit subset is created by a multi-stage automatic pipeline shown in Fig 1 that mimics template-based design. Given a product cutout on a white background and its category, we first retrieve a compatible scene from a predefined background library. The cutout and selected background are then jointly fed into Flux to generate a composed image with the subject placed in context. Finally, we render Chinese marketing copy into predefined text templates according to handcrafted layout rules.

Professional designs. The professional subset contains posters manually crafted by experienced e-commerce designers using standard design software.

For each source we compute the mean scores on Overall, Background, Object, Text, and Layout to characterise its quality profile; the statistics are reported in Fig 2.

Figure 1. **AI-edited posters via Flux.** Given a product cutout and its category (left), we retrieve a matching scene from a predefined background library (middle) and feed both into Flux to compose a subject–background image. The final poster (right) is obtained by adding Chinese marketing copy using predefined text templates.

Table 1. **Train/val/test splits of E-comIQ-18k.**

Source	Train	Val	Test
Merchant HQ	4166	555	279
Merchant LQ	4166	555	279
Open-Source	2500	333	167
AI-generated	1666	222	112
AI-edited	1666	222	112
Prof. design	833	111	56

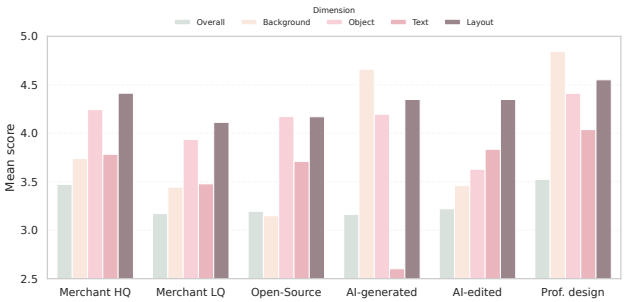

Figure 2. **Mean expert scores by source on E-comIQ-18k.**

A.2. Annotation Checklist and Tag Taxonomy

Table 3 lists the checklist used in E-comIQ-18k. For each image, experts annotate four dimensions (Background, Object, Text, and Layout) with multi-label issue tags and a continuous score in $[1.0, 5.0]$ (one decimal allowed). Tags mark specific defects, while scores summarise the perceived quality of that dimension. A key design choice is the separation between Object and Text. All textual elements printed on the product itself (e.g., brand names and packaging copy) are treated as part of the Object: blurry or malformed packaging text is annotated under Object and only affects the Object score. The Text dimension covers only overlaid marketing copy (titles, slogans, prices, callouts, etc.), where issues such as incorrect line breaks, irrelevant or redundant content, stroke rendering errors, missing text, overlap, or inappropriate font size are recorded. Background and Layout tags focus on global presentation (scene suitability, clutter, balance, occlusion), and an Overall score summarises the commercial usability of the poster given all dimensions.

A.3. Annotation Interface and Reliability

For CoT rationales, we provide a dedicated human-AI collaboration view. Given the expert scores, tags, and image, Qwen-2.5-VL-Max first generates a rationale draft. In the interface, the image is shown on the left, and the model-generated paragraph is shown on the right. Annotators edit the text in a span-based NER-style manner: they can highlight spans to delete, replace them with corrected wording, or insert short additions where the explanation is incomplete; sentences that are entirely incorrect are simply struck out. All edits are recorded, and we compute a character-level edit rate in Chinese, with an average of 32.3% and a maximum of 83.39%, indicating that substantial human refinement is often required.

We further examine reliability across different quality tiers. As shown in Table 2, annotator agreement is highest in the low-quality (1.0–3.0) and high-quality (4.0–5.0) ranges, while the mid-range (3.0–4.0) posters exhibit slightly lower consistency, reflecting the inherent ambiguity of borderline cases. Nevertheless, all tiers maintain reasonably strong agreement, indicating that experts share a stable notion of quality for both clearly good and clearly bad posters.

Table 2. **Inter-annotator stability across quality tiers.** Tiers follow the score ranges (1–3 / 3–4 / 4–5) used in the main paper. Agreement rate is defined as the proportion of samples where annotators' scores differ by no more than 0.5.

Quality Tier (Score Range)	Agreement Rate (%)
Poor (1–3)	74.1
Good (3–4)	62.7
Excellent (4–5)	78.4

Table 3. **Annotation checklist and tag taxonomy.**

Dimension / Issue tags
Background
☐ Color clash with product or brand;
☐ weak scene or context;
☐ irrelevant scene;
☐ cluttered or noisy background;
☐ strong "AI-generated" artefacts;
☐ missing or broken body parts;
☐ heavy cut-and-paste / compositing artefacts;
Other tags: _____;
Score: _____
Object
☐ Illegible or blurry text on the product packaging;
☐ incomplete object contour (parts missing or cut off);
☐ extra or duplicated parts (contour overgrowth);
☐ physically implausible placement or pose;
☐ lighting or perspective inconsistent with the scene;
☐ unreasonable scale or proportion;
☐ visible compositing artefacts;
Other tags: _____;
Score: _____
Text
☐ Incorrect or awkward line breaks;
☐ content irrelevant to the product or promotion;
☐ style mismatch with brand or poster tone;
☐ stroke rendering errors;
☐ spelling mistakes or typos;
☐ missing expected overlaid text;
☐ font too large;
☐ font too small;
☐ overlapping text (with other text or the object);
☐ redundant or repetitive text;
Other tags: _____;
Score: _____
Layout
☐ Overly crowded or cluttered layout;
☐ excessive empty space;
☐ visually unbalanced composition;
☐ important elements occluded or mutually blocking;
Other tags: _____;
Score: _____
Overall
Score: _____

Figure 3. **Examples and prompt for the AI-generated subset.** For each product we take the original category, Chinese title, and a white-background cutout (left), and use GPT-4o with the prompt template on the right to generate a complete e-commerce poster (middle). The prompt enforces strict subject preservation, automatic scene design, and Chinese selling-point copy, so that the generated posters are photorealistic and commercially usable.

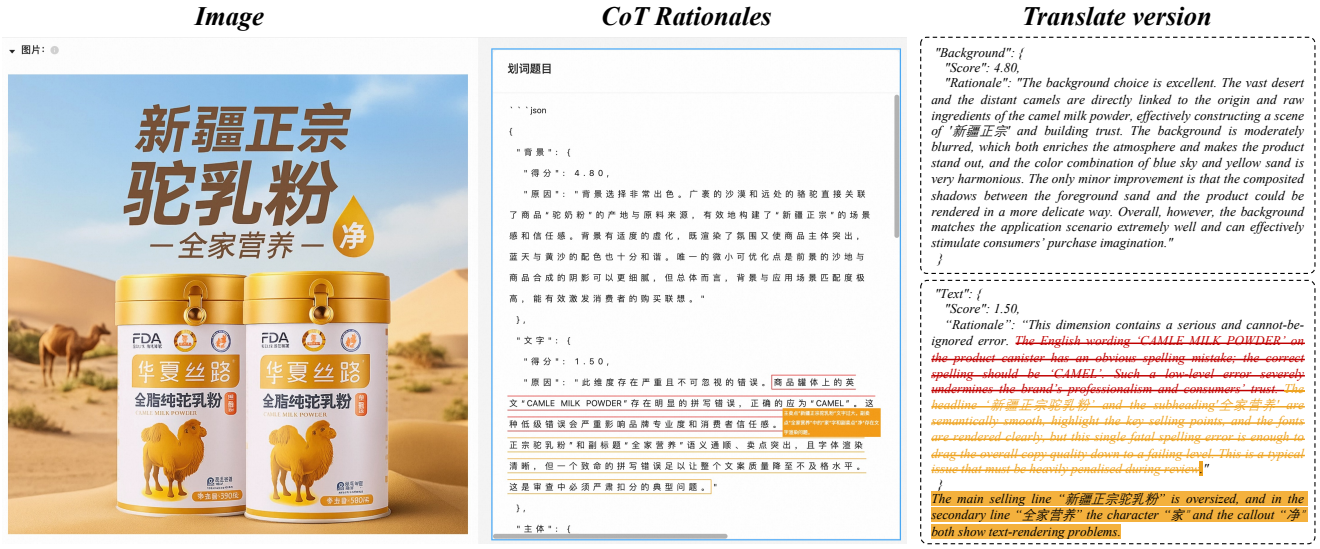

Figure 4. **Example of our CoT editing interface.** Given a poster image (left), the annotator reviews the LLM-generated Chinese rationale (middle) and performs span-level edits to correct errors (shown in red/orange), producing an English translated version (right) that remains faithful to expert judgement.

B. E-comIQ-M: Model and Training Details

B.1. Model Configuration

We build E-comIQ-M on the Qwen2.5-VL-7B-Instruct, using the official vision encoder and tokenizer without architectural changes.

Each sample contains a single poster image and an evaluation instruction. The images are decoded as RGB

(a) Chinese CoT Word Cloud **(b) English CoT Word Cloud**

Figure 5. **Word frequency analysis of model reasoning traces.** We visualize the top frequent words appearing in the model's chain-of-thought (CoT) across 18k e-commerce samples. The Chinese word cloud (left) is directly computed from the original CoT supervision, while the English version (right) is obtained by carefully mapping the top 200 Chinese terms to semantically aligned English phrases. Both visualizations reflect consistent semantic emphasis on background, object clarity, textual quality, composition, and visual communication.

and resized to 512×512 before being fed into the model. Both SFT and GRPO use the same instruction-following format. Given the instruction, the model first produces a Chain-of-Thought in natural-language inside a `<think></think>` block and then outputs a JSON object inside a `<answer></answer>` block.

During training and evaluation, any sample whose `<answer></answer>` block cannot be parsed as valid JSON is treated as invalid , and at inference time we retry decoding up to three times ($temperature = 1.0$, $top\_p = 0.95$, maximum 4096 new tokens, single-turn generation).

B.2. SFT Training Hyperparameters

We perform supervised fine-tuning on the 15k training images using the instruction format in Fig. 6. We use AdamW with a learning rate of 5×10^{-5} and a cosine schedule without warm-up. All parameters of Qwen-2.5-VL-7B-Instruct are updated (full fine-tuning) within the LLaMA-Factory framework. SFT is run with batch size per-GPU $= 1$ and gradient accumulation $= 8$, giving 64 samples per optimizer update. We train for 3 epochs using DeepSpeed ZeRO stage 3 with bf16 precision and a maximum sequence length of 4096 tokens; when the CoT exceeds this limit, only the tail of the reasoning is truncated, while the JSON answer is always kept intact.

B.3. GRPO Training Details

For the second stage we apply GRPO on a hard subset $\mathcal{D}_{\text{hard}}$ of 3k training samples. We initialise the policy π_θ from the SFT checkpoint and use the same model as the frozen reference π_{ref}. The instruction format and target JSON schema are identical to SFT (Fig. 6). For each prompt we sample a group of $G = 4$ continuations with $temperature = 1.0$, $top\_p = 0.95$ and $max\_new\_tokens = 4096$. Invalid JSON outputs are re-sampled up to three times; if all attempts fail, the sample is marked invalid and its reward is set to zero.

Table 4. **Effect of hard-subset size on GRPO performance.** Performance is reported as PLCC and SRCC on the validation split. The default configuration (3k) is highlighted.

Metric	1k	2k	3k (default)	4k	5k
PLCC	0.412	0.421	0.425	0.426	0.423
SRCC	0.423	0.429	0.433	0.431	0.430

Training is carried out on the 3k hard samples for 500 optimizer steps. At each step we process one prompt per device with group size $G = 4$ and use gradient accumulation $= 8$, which yields 32 trajectories per update. We use AdamW with a cosine learning rate schedule, base learning rate 1×10^{-6} and warmup ratio 0.1. The KL regularisation coefficient β in GRPO is set to 0.1. As shown in our logs, the training remains stable without reward collapse.

Hard subset construction. We select $\mathcal{D}_{\text{hard}}$ from the 15k SFT training set using the SFT model's regression error while preserving the source distribution. For each sample j we denote the expert scores by $\mathbf{y}_j \in \mathbb{R}^5$ and the SFT prediction by $\hat{\mathbf{y}}_j \in \mathbb{R}^5$ (four dimensions plus overall). We first compute the mean squared error $\text{MSE}_j = \|\mathbf{y}_j - \hat{\mathbf{y}}_j\|_2^2$, then

SYSTEM PROMPT

As an expert in e-commerce image evaluation and auditing, when a user poses a question about image quality, you must first conduct a thorough reasoning and analysis process before providing the final answer. Both the reasoning process and the final answer must be enclosed in their respective specified tags: wrap the reasoning process with <t></t> tags, and wrap the final answer with <a> tags, as requested by the user.

USER PROMPT

*<image>\nConduct a rigorous diagnostic evaluation of this product image's quality. Assess it in detail across the following four dimensions:\n\n1. **Background and Scene Consistency**:\n - Examine if the background is closely related to the product's actual use case, requiring relevant decorative elements that enhance the sense of a real-world scene and effectively stimulate consumer desire.\n - The background should be clean and uncluttered to emphasize the main product.\n - Strictly check for any absence of product-related elements or the presence of irrelevant elements that could disrupt consumer associations.\n\n2. **Copywriting and Text Quality**:\n - The marketing copy on the product image must be semantically coherent, clearly expressed, highlight key selling points, and be persuasive.\n - Ensure there are no typographical errors, missing words, improper sentence breaks, or disorganized word order in the copy.\n - Pay special attention to subtle rendering errors such as incomplete strokes, character adhesion, blurry fonts, or misspellings during text rendering.\n\n3. **Object and Information Clarity**:\n - Rigorously inspect whether the product itself is complete and clear. Any phenomena such as blurriness, being out of focus, common AI-generated artifacts (e.g., deformation, ghosting, partial blurring, distortion, rough edges, unnatural material blending) are considered severe issues and will result in significant point deductions.\n - The main object must not have indistinct contours, excessive pixel noise, blurring distortion, partial disappearance, misplacement, or abnormal fusion, which are common issues in AI-generated images.\n - Check if key information on the product packaging, such as text and logos, is clear and legible, without any blurring, damage, errors, omissions, or rendering flaws.\n\n4. **Layout and Composition**:\n - Evaluate whether the overall layout is harmonious and aesthetically pleasing, with a clear hierarchy between the main product, text, and background, and no issues of occlusion, crowding, or chaotic arrangement.\n - The placement and size of the text and the main object should be reasonable and not detract from the focus on the product.\n - Assess the use of white space, alignment, visual center, and compositional balance, avoiding visual interference, imbalance, or elements obstructing one another.\n - Strictly inspect for layout flaws such as awkward stitching, element stacking, misalignment, or inefficient use of space.\n\nBased on the overall performance, provide a composite score (out of 5 points), where 0-2 indicates obvious problems, 2-4 indicates minor flaws, and 4-5 indicates excellence in all aspects. Generate your response strictly in the following JSON format: {\"background_score\": [score], \"text_score\": [score], \"object_score\": [score], \"layout_score\": [score], \"total_score\": [score]}. The response format must be: <t>reasoning process</t><a>final answer*

Figure 6. **Instruction and prompt template for E-comIQ-M.** We use a fixed system prompt and a fixed user prompt that ask the model to first provide a `<think>` Chain-of-Thought and then output a JSON object with five scores in the `<answer>` block.

rank samples within each source by MSE_j and take the top fraction so that the final hard subset contains 3k examples with a source mix matching the original 15k set. To verify that our results are not overly sensitive to this choice, we also vary the hard-subset size from 1k to 5k and re-train GRPO. As shown in Table 4, overall PLCC and SRCC on the validation set remain stable across different sizes, with the 3k configuration achieving a slightly better balance between performance and computational cost. We therefore use 3k as the default setting in all main experiments.

Algorithm 1 Construction of hard subset $\mathcal{D}_{\text{hard}}$ for GRPO

Input: Training set $\mathcal{D}_{\text{train}} = \{(x_j, \mathbf{y}_j, s_j)\}_{j=1}^{N}$,
SFT model f_{SFT},
target size K (here $K = 3000$)
Output: Hard subset $\mathcal{D}_{\text{hard}}$

Constants:
  Number of dimensions $D = 5$ (4 sub-scores + overall)
  Source set $\mathcal{S} = \{\text{HQ}, \text{LQ}, \text{Open-Source}, \text{AI-gen.}, \text{AI-edit}, \text{Prof.}\}$
Initialize per-source container $\mathcal{B}_s \leftarrow [\,]$ for all $s \in \mathcal{S}$
for $j = 1$ to N **do**
  $\hat{\mathbf{y}}_j \leftarrow f_{\text{SFT}}(x_j)$ {SFT prediction}
  $e_j \leftarrow \frac{1}{D}\|\hat{\mathbf{y}}_j - \mathbf{y}_j\|_2^2$ {mean squared error}
  Append $(x_j, \mathbf{y}_j, e_j)$ to $\mathcal{B}_{s_j}$
end for
Compute per-source sizes $N_s \leftarrow |\mathcal{B}_s|$ for all $s \in \mathcal{S}$
Set $K_s \leftarrow \left\lfloor K \cdot \frac{N_s}{\sum_{s' \in \mathcal{S}} N_{s'}} \right\rfloor$ for all $s \in \mathcal{S}$
Initialize $\mathcal{D}_{\text{hard}} \leftarrow \emptyset$
for each source $s \in \mathcal{S}$ **do**
  Sort $\mathcal{B}_s$ in **descending** order of e_j
  Take the first K_s samples from $\mathcal{B}_s$ and add them to $\mathcal{D}_{\text{hard}}$
end for
return $\mathcal{D}_{\text{hard}}$

B.4. Reward Design and Ablations

As discussed in Sec. 4.2 and Table 5, combining the accuracy and distribution terms ($R_{\text{acc}} + R_{\text{dist}}$) on top of SFT gives the best overall performance. Here we provide additional analysis on the reward weights and GRPO optimisation hyperparameters.

Figure 7. Additional qualitative examples.

Reward weight sensitivity. Figure 8 studies the effect of the accuracy tolerance τ in R_{acc}. Very loose thresholds ($\tau \geq 0.5$) make the reward less informative and lead to weaker PLCC and SRCC, while very strict thresholds ($\tau = 0.1$) also hurt performance. The curves are most stable around $\tau = 0.2$, which we adopt as the default. Figure 9 varies the accuracy weight λ_{score} that balances R_{acc} and R_{dist}. Using only the accuracy term ($\lambda_{\text{score}} = 1.0$) or giving it too little weight ($\lambda_{\text{score}} \leq 0.45$) degrades both correlations. The best trade off is obtained near $\lambda_{\text{score}} = 0.65$, confirming that a moderate contribution from the distribution term helps align the geometry of sub scores with expert ratings.

GRPO optimisation hyperparameters. We further sweep the GRPO learning rate and KL penalty coefficient β (see Table 6). Very small learning rates or β values slow down optimisation and

Table 5. **Ablation on Q-Insight.** PLCC/SRCC (top rows) and Acc@0.5/1.0 (bottom rows). **Exp1**: SFT; **Exp2**: SFT+GRPO.

Model	Overall	Background	Subject	Text	Layout
Exp1	0.297/0.319	0.442/0.478	0.242/0.244	0.291/0.304	0.379/0.391
Exp2	0.338/0.348	0.459/0.496	0.386/0.304	0.320/0.342	0.375/0.403
Exp1	50.8/75.2	63.0/81.6	49.4/74.4	43.8/67.4	55.2/77.8
Exp2	53.6/79.6	60.2/78.4	51.4/74.0	47.2/68.4	57.8/77.2

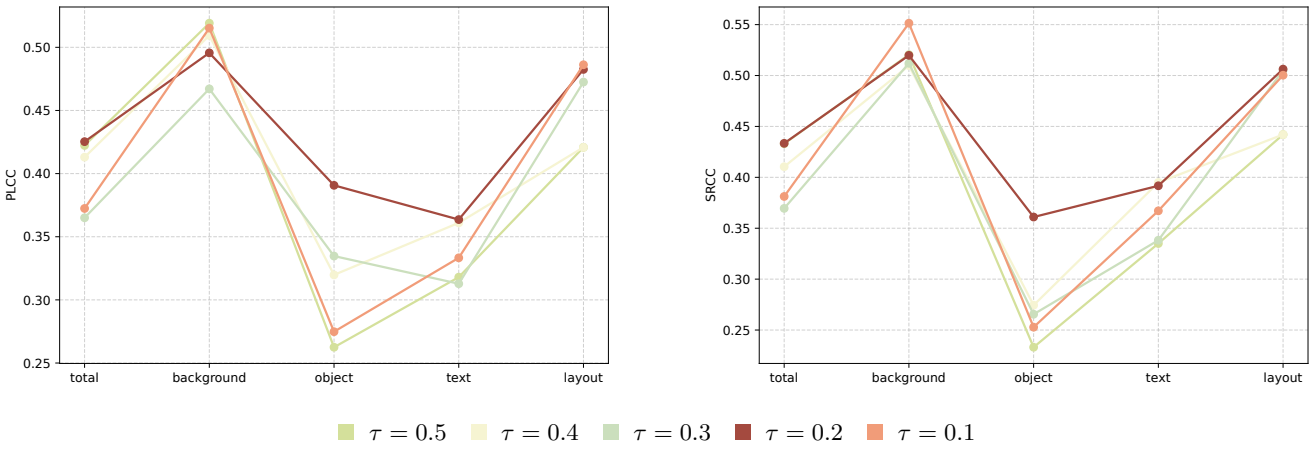

Figure 8. **Effect of the accuracy tolerance τ on reward performance.** Left: overall PLCC across the four sub-dimensions (Background, Object, Text, Layout) and the total score under different τ values. Right: corresponding SRCC results. Each coloured line denotes a different tolerance setting.

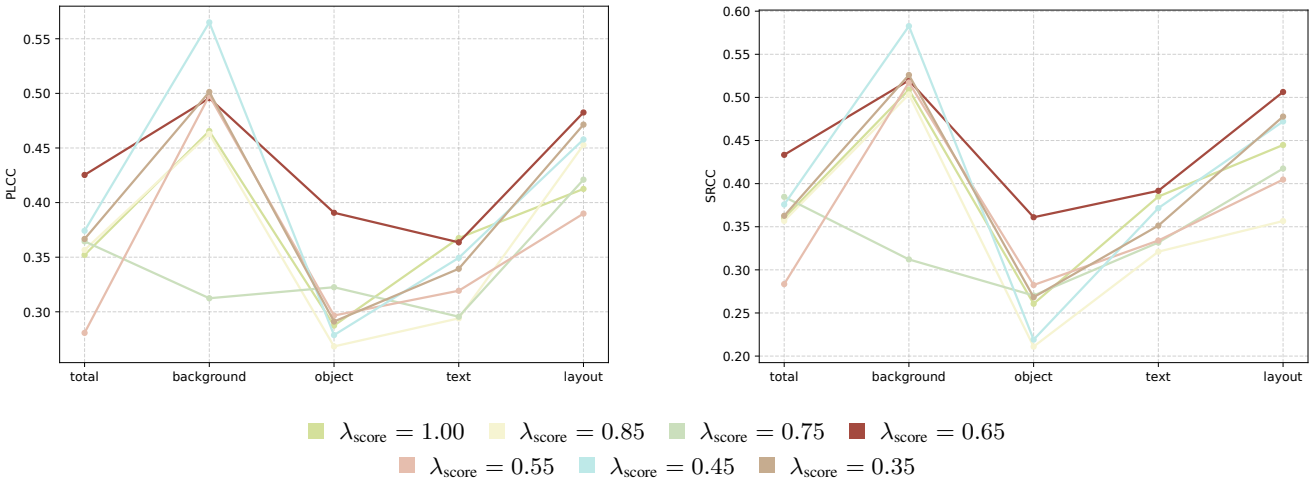

Figure 9. **Effect of the accuracy weight λ_{score} on reward performance.** Left: PLCC across the four sub-dimensions (Background, Object, Text, Layout) and the total score under different λ_{score} values. Right: corresponding SRCC results. Each coloured line denotes a different setting of the accuracy weight in the reward.

yield limited gains over SFT, while larger values lead to unstable training and a drop in PLCC/SRCC. The final setting used in the main experiments ($\text{lr} = 1 \times 10^{-6}$, $\beta = 0.1$) lies in a stable region and offers the best overall balance between convergence speed and evaluation performance.

C. E-comIQ-Bench and Evaluation Toolbox

C.1. Prompt Design and Generation Setup

The construction procedure of E-comIQ-Bench follows the design described in Sec. 5.1. Each case contains a foreground cutout,

its merchant poster, and a Chinese prompt derived from the product's selling points. Here we provide additional implementation details regarding prompt generation and image synthesis.

Prompt generation. Building on Sec. 5.1, we now detail how the Chinese poster prompts are constructed. Given a product cutout, the original merchant poster, and structured product attributes (category, short title, and selling points), we first query Qwen2.5-VL-72B to generate a high-level poster prompt. To reduce prompt bias, we design five template variants covering different copywriting styles, typography rules, and layout strategies. One template is randomly selected and filled with the extracted attributes, and the model rewrites the text with improved commercial tone and layout instructions. The output is a complete "generation prompt" that will later be used to produce the final poster image.

Importantly, the template itself is *not* used for generation: it only guides the rewritings made by Qwen2.5-VL-72B. The rewritten result becomes the actual prompt supplied to different text-to-image systems (Fig. 10). Examples from seven categories are

Table 6. **Effect of GRPO learning rate and KL coefficient on correlation performance.** Each cell reports **PLCC / SRCC** on the E-comIQ-18k test set.

Setting		Overall		Background		Object		Text		Layout	
lr	β	*PLCC*	*SRCC*	*PLCC*	*SRCC*	*PLCC*	*SRCC*	*PLCC*	*SRCC*	*PLCC*	*SRCC*
1×10^{-6}	0.10	**0.425**	**0.433**	**0.496**	0.520	**0.391**	**0.361**	**0.364**	**0.392**	**0.483**	**0.506**
5×10^{-5}	0.10	0.267	0.277	0.383	0.462	0.189	0.223	0.270	0.270	0.375	0.403
1×10^{-6}	0.05	0.329	0.348	0.441	0.508	0.263	0.239	0.342	0.346	0.431	0.450
5×10^{-5}	0.05	0.326	0.302	0.417	0.454	0.265	0.237	0.226	0.311	0.455	0.488
1×10^{-6}	0	0.346	0.346	0.458	**0.530**	0.272	0.238	0.272	0.283	0.390	0.418
5×10^{-5}	0	0.265	0.235	0.312	0.312	0.123	0.070	0.096	0.132	0.221	0.218

Template 1

USER PROMPT

*Generate a single, master-level **English text-to-image prompt**. This prompt must guide an AI to create a flawless, 800*800, 1:1, high-converting e-commerce product image.*
Your generated prompt MUST ensure the final image perfectly executes these Four Core Principles:
1. Background & Scene Relevancy:
- The background must evoke the feeling and use-case of the product ('{category}').
*- Include tasteful, relevant props/elements that enhance the scene, but maintain a clean, premium aesthetic. **Avoid all distractions.***
2. Master-Level Copywriting & Typographic Artistry (CRITICAL REFINEMENT):
Your process is now more refined. You will decide on content and style in separate, deliberate steps.
Step A: Critical Analysis.
- Deeply analyze all inputs: the product, `short_title` ('{short_title}'), `selling_points` ('{selling_points}'), and crucially, the reference 'original image'.
Step B: Core Copywriting Decision (Content).
*Make a judgment on the **text content** by choosing one of two paths:*
*- **PATH 1 (Adopt Content):** IF the original image's text is strong, your prompt will use the **exact Chinese text** from it.*
*- **PATH 2 (Innovate Content):** IF the original text is weak or absent, **creatively generate new Chinese text** (4-20 characters) using advanced structures (Poetic Slogan, Dynamic Headline/Tagline).*
Step C: Typographic Styling & Enhancement (Style - YOUR NEW KEY FOCUS).
This is where your artistry shines. After deciding the text content in Step B, you will now design its visual style.
*- **1. Dissect the Style DNA:** First, **meticulously analyze the visual style of the text in the original image**. Your prompt's description must be based on this analysis. Consider:*
*- **Font Family:** Is it a Serif (衬线体), Sans-serif (无衬线体), Script (手写体)?*
*- **Weight & Form:** Is it Bold (粗体), Light (细体), Condensed (窄体), Italic (斜体)?*
*- **Effects & Rendering:** Does it have a Drop Shadow (投影), Outer Glow (外发光), Gradient (渐变), Emboss (浮雕), Outline (描边)?*
*- **Color Palette:** What are the primary text colors?*
*- **2. Choose the Styling Strategy:** Based on your analysis, your prompt will describe one of two styling strategies:*
*- **Strategy A (High-Fidelity Enhancement):** If the original style is good, your prompt will describe a **premium, high-fidelity recreation** of it. This means using the same core style but making it sharper, cleaner, and more perfectly rendered. For example, your prompt might say, "recreating the original's bold sans-serif style but with flawless vector-quality rendering and a more subtle, realistic soft drop shadow."*
*- **Strategy B (Creative Evolution):** If the original style is basic or can be improved, your prompt will describe a **creative evolution inspired by the original**. This means retaining the original's spirit but elevating it. For example, your prompt might say, "evolving the original's simple white text by rendering it in a modern, light-weight sans-serif font, and adding a gentle, clean neon glow that matches the product's high-tech aesthetic."*
*- **3. Specify Dynamic Layout:** Even when adopting a style, ensure the layout is dynamic. Your prompt should describe a sophisticated arrangement, emphasizing **hierarchy through varied font sizes and weights**, and using intelligent placement (e.g., "a large impactful headline with a smaller tagline nestled below it").*
Step D: Universal Quality Mandate.
*Regardless of the path taken, the prompt MUST demand **Zero-Tolerance Quality Control**. Use phrases like: 'Flawless, vector-quality Chinese typography with perfect anti-aliasing. Absolutely no artifacts, pixelation, missing strokes, smudging, or character distortion.'*
3. Product Subject & Information Clarity.
*- The product must be depicted with **hyper-realistic detail and perfect lighting**.*
*- Your prompt must explicitly forbid AI flaws: **no blur, distortion, warping, ghosting, unrealistic textures, or unnatural blending.***
4. Overall Composition & Balance:
- The prompt must describe a composition that is balanced, harmonious, and professional, with ample negative space.
Final Output Requirement:
*- **Output ONLY the final, complete English prompt text.**
- Do not include any commentary. Be direct and professional.*

Template 2

USER PROMPT

*Generate a single, master-level **English text-to-image prompt**. This prompt will be used by an AI to create a stunning, 800*800,1:1,high-converting e-commerce product image.*
Your generated prompt MUST ensure the final image perfectly executes these Four Core Principles:
1. Background & Scene Relevancy:
- The background must evoke the feeling and use-case of the product ('{category}').
*- Include tasteful, relevant props/elements that enhance the scene, but maintain a clean, premium aesthetic. **Avoid all distractions.***
2. Master-Level Copywriting & Typographic Artistry (WITH CONDITIONAL LOGIC):
This is your most critical task. You must first evaluate the provided materials and then choose the best creative path.
Step A: Critical Analysis & Evaluation.
*- Deeply analyze all inputs: the product itself, the `short_title` ('{short_title}'), the core `selling_points` ('{selling_points}'), and most importantly, **the text content, layout, and design quality within the reference 'original image'.***
Step B: The Core Creative Decision & Action.
You must now make a critical judgment and choose one of the following two paths:
--- PATH 1 (PRIORITY): Adopt & Refine ---
***Choose this path IF** the reference 'original image' already contains well-designed, clear, and compelling Chinese selling points. Your main task is to leverage this existing success. Your generated prompt will then describe:*
*- **1. Text Content:** Explicitly state that the prompt should use the **exact Chinese text from the original image**.*
*- **2. Layout & Composition.** Describe a layout that **respects and enhances the original composition**. Your prompt should guide the AI to perfect the alignment, spacing, and visual balance of the existing text layout.*
*- **3. Rendering Quality:** Specify font styles and supreme rendering quality to make the *existing* text look even more premium and flawless than in the reference.*
--- PATH 2 (FALLBACK): Create & Innovate ---
***Choose this path IF** the reference image has minimal, poor, badly rendered, or no text at all. You must then take full creative control. Your generated prompt will describe:*
*- **1. Text Content:** **Creatively generate new Chinese text** (total 4-20 characters) by synthesizing all inputs. Use advanced structures like a Poetic Slogan or a Dynamic Headline/Tagline.*
*- **2. Layout & Composition.** Design a **completely new, dynamic typographic layout**. Use concepts like font size interplay, vertical stacking, perspective integration, or text wrapping to create visual excitement.*
*- **3. Font & Style:** Apply **Font Psychology**, choosing a font that matches the product's personality (e.g., 'clean sans-serif for tech', 'elegant serif for luxury') and describe it clearly.*
Step C: Universal Quality Mandate.
***Regardless of which path you take**, the prompt MUST demand **Zero-Tolerance Quality Control** for all text. Use phrases like: 'Flawless, vector-quality Chinese typography with perfect anti-aliasing. Absolutely no artifacts, pixelation, missing strokes, smudging, or character distortion.'*
3. Product Subject & Information Clarity:
*- The product must be depicted with **hyper-realistic detail and perfect lighting**.*
*- Your prompt must explicitly forbid AI flaws: **no blur, distortion, warping, ghosting, unrealistic textures, or unnatural blending.***
4. Overall Composition & Balance:
- The prompt must describe a composition that is balanced, harmonious, and professional.
- If Path 1 was chosen, this means perfecting the original's balance. If Path 2, it means creating a new, compelling balance. Ensure ample negative space.
Final Output Requirement:
*- **Output ONLY the final, complete English prompt text.**
- Do not include any commentary. Be direct and professional.
- Your final prompt should be a masterpiece of detail and creative direction, reflecting your chosen path.*

Figure 10. **Prompt templates for generating Chinese e-commerce poster instructions.** We provide the white-background cutout, the original merchant poster, and structured attributes (category, short title, and selling points) to Qwen2.5-VL-72B. The model rewrites the information into a professional, stylistic *generation prompt* that follows detailed typography, layout, and text-quality constraints. Five templates are used to encourage stylistic diversity.

shown in Fig. 11 and Fig. 12.

Generation setup. All text-to-image models are queried with a unified image resolution of 800 × 800 pixels, and all models support both Chinese and English prompts without requiring language-specific templates. We evaluate a mixture of commercial and open-source systems, including Seedream 4.0 [6], GPT-4o [**?**], Gemini-2.5-Flash-Image [1], Flux-Kontext-max [3], and the open-source Qwen-Image-Edit [7]. Commercial models are accessed through official HTTP APIs, whereas Qwen-Image-Edit is executed locally via the official SDK. For all models, we follow their default inference settings (e.g., sampling steps and classifier-free guidance), because our benchmark focuses on cross-model stylistic controllability rather than model-specific tuning. A unified robustness policy is used across systems: if one query fails due to network or decoding errors, we automatically retry the same request up to three times. Only successful generations are kept as final benchmark samples.

C.2. Automatic Evaluation Toolbox

E-comIQ-Bench evaluates each generated poster along the same four quality dimensions as human annotation (Background, Object, Text, Layout) plus the overall score. Human evaluation and our proposed E-comIQ-M serve as the two reference metrics. In addition, two auxiliary indicators are used to measure structural correctness: object consistency and text accuracy.

Object consistency. Given a cutout, we measure how well the

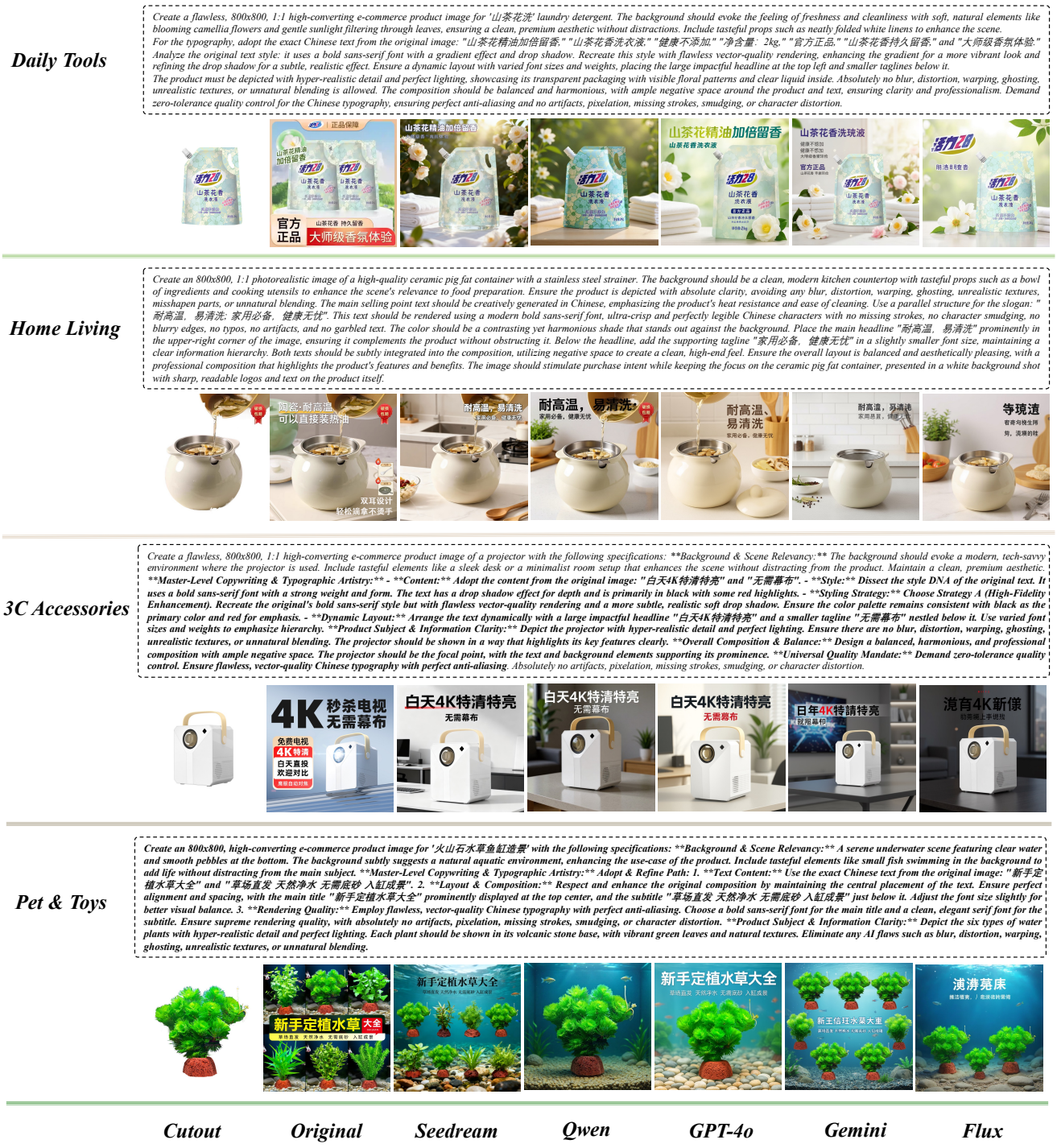

Figure 11. **Prompt–generation showcase across product categories (1/2).** For each cutout–prompt pair, several commercial systems and research models are queried to generate one poster per model. The merchant poster is shown as the human-designed reference. Categories on this page include *Daily Tools, Home Living*, and *3C Accessories*.

generated poster preserves the true object identity. We first use DINOv2 [5] to detect the target category in the poster and obtain the corresponding bounding region, then apply SAM-HQ [2] followed by Vitmatte [8] to obtain a refined object mask. The extracted object region is compared against the original cutout using DINO feature cosine similarity, CLIP image embedding cosine similar-

Figure 12. **Prompt–generation showcase across product categories (2/2).** The same protocol applies to the remaining categories: *Fresh Food, Daily Tools (additional cases), Footwear & Apparel*, and *Pet & Toys*.

ity, and LPIPS perceptual distance. These metrics quantify semantic consistency (DINO/CLIP) and pixel-level fidelity (LPIPS) between the generated image and the ground-truth product.

Text accuracy. Unlike object text on the product packaging (assessed in the Object dimension), this metric evaluates whether the generated marketing copy faithfully reflects the intended prompt semantics. We train a lightweight text extractor to obtain structured selling-point keywords from each poster, and compare them to the prompt using two levels of textual matching: (1) Sentence-level structural accuracy measured by F1 over detected key phrases (Phrase F1). (2) Character-level normalised Levenshtein similarity computed as the Bag-of-Characters cosine similarity (Char Sim), which ignores ordering but enforces semantic token agreement. This combination penalises both missing key information and hallucinated claims.

Reproducibility. All metrics are implemented in our evaluation toolbox, which will be released together with the dataset. To ensure fairness and stability, the toolbox retries API failures up to three times and outputs merged JSON statistics per model. The snippet below illustrates the text-matching aggregation used in the benchmark.

References

[1] Alisa Fortin, Guillaume Vernade, Kat Kampf, and Ammaar Reshi. Introducing gemini 2.5 flash image, our state-of-the-art image model. *Google Developers Blog*, 26, 2025. 7

[2] Lei Ke, Mingqiao Ye, Martin Danelljan, Yu-Wing Tai, Chi-

Keung Tang, Fisher Yu, et al. Segment anything in high quality. *Advances in Neural Information Processing Systems*, 36: 29914–29934, 2023. 8

[3] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. 7

[4] Jinpeng Lin, Min Zhou, Ye Ma, Yifan Gao, Chenxi Fei, Yangjian Chen, Zhang Yu, and Tiezheng Ge. Autoposter: A highly automatic and content-aware design system for advertising poster generation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1250–1260, 2023. 1

[5] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research Journal*, 2024. 8

[6] 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. 7

[7] 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. 7

[8] Jingfeng Yao, Xinggang Wang, Shusheng Yang, and Baoyuan Wang. Vitmatte: Boosting image matting with pre-trained plain vision transformers. *Information Fusion*, 103:102091, 2024. 8