FIDPOSTER: END-TO-END PRODUCT POSTER GENERATION WITH FINE-GRAINED PRODUCT DETAIL FIDELITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Fine-grained Product Detail Fidelity refers to preserving a reference product's packaging text, textures, and colors in generated posters. Existing models often distort these details, compromising commercial usability despite visually appealing designs. Our analysis shows that VAE reconstruction degrades fine-grained details, highlighting the need for direct supervision of decoded appearance. During sampling, the absence of explicit fidelity guidance can further allow product details to drift from the reference. These observations motivate **FidPoster**, an end-to-end framework for **Fid**elity-preserving product **Poster** generation that combines pixel-space fidelity learning with reference-guided trajectory correction. During training, **Pixel-Space Fidelity Activation (PFA)** directly supervises decoded product regions to align optimization with visible detail preservation. During inference, **Energy-Guided Velocity Correction (EVC)** uses bidirectional correspondence energy to measure inconsistent and missing details relative to the reference. Gradients of this energy correct the sampling velocity to suppress detail drift without requiring externally provided product masks at inference. We also construct **FidPoster14K** with 14K product-poster pairs and **FidPoster-Bench**, a 1,000-sample benchmark comprising an 800-sample *Fine-grained Fidelity* setting for controlled fidelity evaluation and a 200-sample *Independent Reference* setting for assessing generalization to independently sourced products. Experiments across both settings demonstrate consistent improvements in product-detail fidelity while maintaining competitive promotional-text accuracy and poster quality.

1 INTRODUCTION

Product poster generation aims to create a complete commercial poster from a reference product image and a textual design instruction. Beyond visual appeal and coherent composition, commercial use requires **Fine-grained Product Detail Fidelity**: preserving distinctive product characteristics, including packaging text, textures, and colors, in the generated poster. Despite substantial progress in poster generation, existing methods Vinaykumar et al. (2026); Cui et al. (2026); Qin et al. (2026); Ma et al. (2025) may preserve overall product appearance while distorting packaging text, numerical specifications, logos, or local structures, as illustrated in Figure 1(a). Such errors can misrepresent the advertised product and render visually appealing posters commercially unusable.

We examine this challenge through an analysis of generation failures and autoencoder reconstruction. Across the same 200 test inputs, product-detail errors occur in 57.5% of posters generated by Qwen-Image-Edit Wu et al. (2025) and 85.0% of those generated by PosterOmni Chen et al. (2026), as shown in Figure 1(c). These results reveal a substantial gap between existing generation capabilities and the fidelity required for commercial product representation. More details are provided in Appendix A. To investigate a potential source of degradation, we examine the autoencoders underlying latent-space generation. Many recent methods Wu et al. (2025); Labs (2025); Seedream et al. (2025); Cai et al. (2025); Team et al. (2026) use a variational autoencoder (VAE) to compress images into latent representations and decode generated latents into images Esser et al. (2024). We isolate compression and reconstruction effects by directly encoding and decoding ground-truth posters with four representative autoencoders Wu et al. (2025); Labs (2025); Labs et al. (2025); Esser et al.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

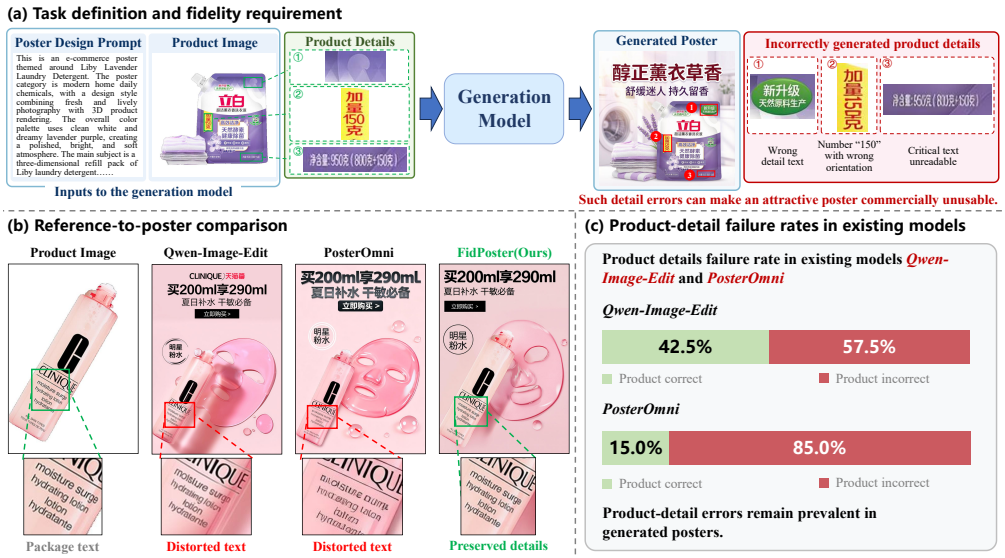


Figure 1: Motivation for fine-grained product-detail fidelity. (a) Existing generation models may distort critical product details despite producing visually plausible posters. (b) FidPoster better preserves packaging text and other local product details. (c) Existing generation models exhibit a high incidence of product-detail distortions.

Table 1: Product-detail degradation under direct autoencoder reconstruction on the *Fine-grained Fidelity* setting of FidPoster-Bench. Red values in parentheses for Avg. Recon. denote changes from the GT reference, with all changes indicating degraded reconstruction fidelity.

Setting	SSIM↑	GMSD↓	Color↓	LF Err.↓	HF Err.↓	HF/LF
GT (Ref.)	1.0000	0.0000	0.0000	0.0000	0.0000	–
Qwen	0.8379	0.1226	3.0705	0.1990	0.7632	3.8344
FLUX.1	0.7989	0.1290	3.7984	0.2904	0.7835	2.6984
FLUX.2	0.8029	0.1247	3.8904	0.2972	0.7642	2.5712
SD3.5	0.7776	0.1381	4.1178	0.3013	0.8671	2.8785
Avg. Recon.	**0.8043** (-0.1957)	**0.1286** (+0.1286)	**3.7193** (+3.7193)	**0.2720** (+0.2720)	**0.7945** (+0.7945)	**2.9956**

(2024), without diffusion sampling or latent prediction. Table 1 shows an average SSIM decrease of 0.1957, GMSD of 0.1286, and color error of 3.7193. The mean high-to-low-frequency reconstruction error ratio reaches 3.00, indicating greater degradation of high-frequency details. These errors arise even when decoding encoded ground-truth latents, showing that matching latent targets does not ensure faithful decoded appearance. This motivates direct supervision of decoded product details, particularly local structures and high-frequency patterns.

Motivated by these findings, we propose **FidPoster**, which combines pixel-space fidelity learning during training with reference-guided trajectory correction during inference. **Pixel-Space Fidelity Activation (PFA)** retains the standard latent-space objective for overall poster generation while adding appearance, edge, and multi-scale structural specifically within decoded product regions. This combines latent-space learning for background synthesis and global composition with direct pixel-space supervision for fine-grained product fidelity. PFA also trains a lightweight product localizer to enable region-specific fidelity guidance during inference. During sampling, reference conditioning alone provides no explicit discrepancy feedback to counter product-detail drift. **Energy-Guided Velocity Correction (EVC)** therefore formulates detail preservation as fidelity-aware trajectory optimization. At each guided step, the trained localizer predicts the product region, restricting fidelity-energy computation to product content and reducing interference from the surrounding background. Within this region, a bidirectional correspondence energy compares decoded

details with the reference: generated-to-reference matching penalizes inconsistent details, while reverse matching penalizes missing details. Gradients of this energy then correct the sampling velocity to suppress detail drift without requiring externally provided product masks at inference.

To support fine-grained product fidelity learning, we construct **FidPoster14K**, a dataset of 14K product–poster pairs curated through product-focused screening and human selection. Each example includes an extracted reference product, a target poster, a product-region mask, and a design instruction, providing appearance correspondence and regional supervision for learning detail preservation. We further introduce **FidPoster-Bench** for systematic evaluation of reference-product fidelity, covering packaging text, local structures, high-frequency details, and color consistency. We will publicly release the datasets, annotations, evaluation toolkit, source code, and model weights.

The main contributions are:

- Fine-grained product-detail fidelity is systematically investigated through generation failure analysis and controlled VAE reconstruction experiments, revealing frequent detail distortions and the limitations of latent-target matching in ensuring decoded fidelity.
- We propose **FidPoster**, combining product-focused pixel-space supervision with reference-guided trajectory optimization. PFA learns to preserve visible details, while EVC uses bidirectional correspondence feedback to suppress sampling drift without requiring externally provided product masks at inference.
- We construct **FidPoster14K** with product-region annotations and design instructions for fidelity learning, and **FidPoster-Bench** for systematic evaluation of packaging text, local structures, high-frequency details, and colors.

2 Related Work

Product Poster Generation. Existing methods broadly adopt modular pipelines or unified generative frameworks. Modular approaches Li et al. (2023); Chen et al. (2025); Liu et al. (2026); Lin et al. (2023); Wang et al. (2024); Fan et al. (2026) organize copywriting, layout planning, and rendering through intermediate design representations, offering explicit control at the cost of pipeline complexity. Diffusion-based methods increasingly integrate scene and text generation Ma et al. (2025); Gao et al. (2025), while DreamPoster Hu et al. (2025) and PosterOmni Chen et al. (2026) advance unified image-conditioned generation through multi-task learning and reward feedback. PosterMaker Gao et al. (2025) requires subject masks and explicit text positions, preventing direct application to our product-image-and-instruction setting without additional preprocessing or layout planning. SimplePoster Cui et al. (2026) simplifies inpainting-based generation through full-parameter fine-tuning and character-level position encoding, emphasizing subject preservation and text placement. FidPoster further targets fine-grained fidelity of packaging text, local structures, and high-frequency details through product-region pixel-space supervision and reference-guided sampling correction, without requiring externally provided product masks at inference or text layouts at inference.

Poster Datasets and Benchmarks. Existing datasets support poster generation through text annotations, product references, and diverse design tasks. POSTER-TEXT Gao et al. (2023) provides sentence-level text and bounding-box annotations for typography design. PosterMaker Gao et al. (2025) includes foreground masks and text annotations for joint scene and text generation, while AutoPP1M Fan et al. (2026) combines large-scale product posters with user feedback for generation and commercial optimization. DreamPoster Hu et al. (2025) constructs source-image–poster pairs with textual and typographic annotations, and PosterOmni-200K Chen et al. (2026) covers six tasks spanning entity-preserving editing and global design. Corresponding benchmarks evaluate text accuracy, subject consistency, instruction following, and aesthetic quality, supporting broad assessment of poster generation. Complementing these resources, FidPoster14K emphasizes clearly identifiable product features and provides appearance correspondence with regional supervision for fine-grained fidelity learning. FidPoster-Bench further evaluates packaging-text fidelity, local structures, high-frequency details, and color consistency, enabling targeted assessment of product-detail preservation alongside overall poster quality.

3 FIDPOSTER14K AND FIDPOSTER-BENCH

Existing poster datasets emphasize scale, typography, or design diversity, but are not explicitly curated for fine-grained product fidelity: product details may be unclear, reference–target correspondence may be ambiguous, and regional annotations are often unavailable. Existing benchmarks also tend to conflate product preservation with overall poster quality. We address these gaps with **FidPoster14K** for fidelity learning and **FidPoster-Bench** for evaluation under Fine-grained Fidelity and Independent Reference settings.

3.1 FIDPOSTER14K DATASET

Starting from approximately 80K posters from AutoPP1M Fan et al. (2026) and POSTER-TEXT Gao et al. (2023), we remove samples with ambiguous or occluded products, insufficient resolution, degraded details, incomplete content, or defective layouts. Only 14,067 samples pass automatic filtering and manual inspection, including 8,971 square posters at 800×800 and 5,096 portrait posters at 513×750. This strict curation concentrates examples in which product details are sufficiently visible for fidelity supervision.

Each sample pairs a target poster with an extracted reference product, a detailed design instruction, and product and promotional-text masks. Unrelated logos and warning text are removed using LaMa Suvorov et al. (2021), and Gemini 2.5 Pro Comanici et al. (2025) describes the poster content and design. Extracting the reference from its target poster provides reliable appearance correspondence, while regional masks enable product details to be supervised separately from the surrounding design. All annotations are manually verified. These properties support decoded pixel-space supervision without simplifying the task into product compositing; inference still requires only a product image and a design instruction. Examples are provided in Appendix B.

3.2 FIDPOSTER-BENCH

FidPoster-Bench contains 1,000 samples under two complementary settings. The *Fine-grained Fidelity* setting contains 800 manually verified product–poster pairs, disjoint from FidPoster14K, and provides target posters and regional annotations for reproducible quantitative evaluation. Within this setting, the partially overlapping *Text-Fidelity* subset contains 400 products with packaging text, while the *Visual-Fidelity* subset contains 529 products with distinctive textures, local structures, or other fine-grained characteristics. These subsets separately expose textual and visual product distortions that may be obscured by global similarity measures. Poster Text Score and FID are additionally evaluated on all 800 samples to verify that product preservation is not achieved at the expense of promotional-copy accuracy or overall poster quality.

The *Independent Reference* setting contains 200 independently sourced product images from Products-10K Bai et al. (2020), paired with design instructions but without target posters. Blind human evaluation on this setting tests whether fidelity improvements transfer beyond same-source product–poster pairs. Together, the two settings distinguish visually appealing but product-inaccurate results from faithful but poorly designed posters, which neither setting can reveal alone. Further details are provided in Appendix C.

4 METHODOLOGY

4.1 OVERVIEW

As illustrated in Figure 2, FidPoster consists of two complementary stages. Pixel-Space Fidelity Activation (PFA) learns *what to preserve* by directly supervising decoded product details, and *where to preserve* through reference-aware product localization. Energy-Guided Velocity Correction (EVC) then performs sample-specific trajectory correction during inference through three steps: (1) *locate the product*, (2) *measure detail deviations* from the reference, and (3) *correct the sampling velocity*. Together, PFA establishes fine-grained preservation capability, while EVC provides reference-specific correction along the sampling trajectory.

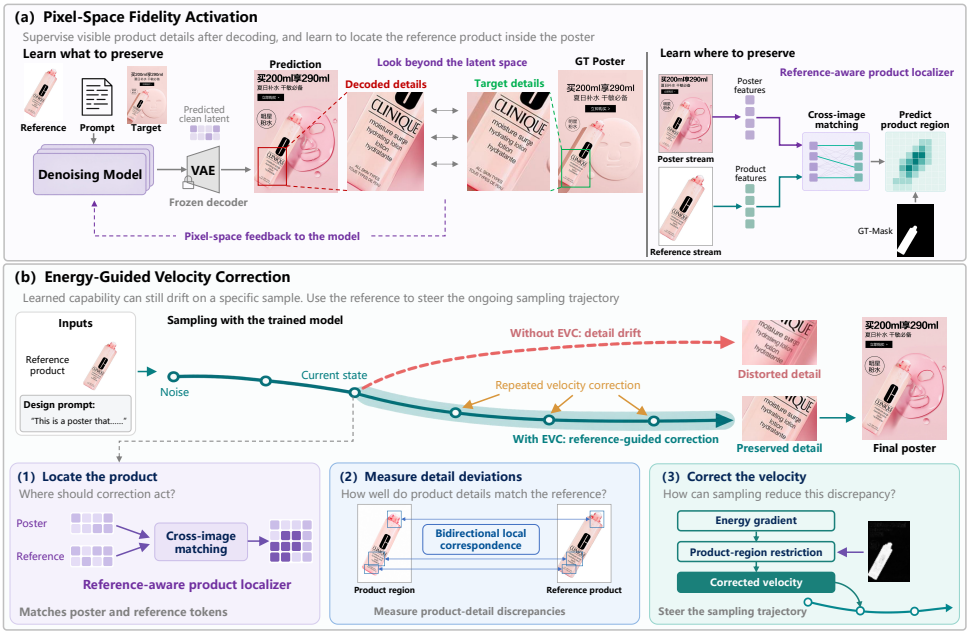


Figure 2: **Overview of FidPoster.** (a) Pixel-Space Fidelity Activation (PFA) directly supervises decoded product details while learning a reference-aware product localizer. (b) Energy-Guided Velocity Correction (EVC) localizes the emerging product, measures bidirectional detail discrepancies against the reference, and repeatedly corrects the sampling velocity to suppress detail drift.

4.2 Pixel-Space Fidelity Activation

The standard flow-matching objective operates in latent space, while fine-grained product fidelity is determined after decoding. As shown in Section 1, latent compression and reconstruction can degrade product details, motivating **Pixel-Space Fidelity Activation (PFA)**. PFA decodes the predicted clean latent during training and directly supervises the product region in pixel space, complementing the original flow-matching objective.

Following the linear flow-matching formulation, we estimate the clean latent and decode it as

$$\hat{\mathbf{z}}_0 = \mathbf{z}_t - t\hat{\mathbf{v}}_t, \qquad \hat{\mathbf{I}}_0 = \mathcal{D}(\hat{\mathbf{z}}_0), \tag{1}$$

where $\mathcal{D}$ denotes the frozen VAE decoder. We directly compare $\hat{\mathbf{I}}_0$ with the original ground-truth poster $\mathbf{I}_{\text{gt}}$, allowing gradients to propagate through the frozen decoder to the denoising model.

To capture complementary aspects of product fidelity, we impose three pixel-space objectives within the ground-truth product region:

$$\mathcal{L}_{\text{pix}} = \lambda_{\text{char}}\mathcal{L}^p_{\text{char}} + \lambda_{\text{sob}}\mathcal{L}^p_{\text{sob}} + \lambda_{\text{str}}\mathcal{L}^p_{\text{str}}. \tag{2}$$

Specifically, the Charbonnier loss constrains local appearance, the Sobel loss emphasizes edges and fine structures, and the masked MS-SSIM loss encourages multi-scale structural consistency. All pixel-space objectives are weighted by $w(t) = (1-t)^\gamma$, assigning stronger supervision to lower-noise states where the clean-image estimate is more reliable.

Reference-aware product localization. The pixel-space objectives exploit the product masks available during training. However, requiring an externally specified product mask at inference time would violate our end-to-end setting. More importantly, Stage II requires the location of the generated product to perform sample-specific fidelity correction. We therefore jointly learn a lightweight reference-aware product localizer using the product-region annotations available during training.

Specifically, the visual Transformer features are separated into poster-stream tokens $\mathbf{H}_g$ and reference-product tokens $\mathbf{H}_r$. The two streams are projected into a shared matching space, where pairwise similarities between poster and reference-product tokens are computed and aggregated

across reference tokens and attention heads to produce a soft product-location map $\hat{\mathbf{M}}_p$. By explicitly matching the poster stream with the reference product, the localizer exploits their cross-image correspondence rather than estimating the product region solely from the poster representation. The detailed matching architecture and mask construction procedure are provided in Appendix D.1.

During training, the ground-truth product mask is mapped to the corresponding poster-token resolution to obtain $\mathbf{M}_p^*$, and the localizer is supervised by

$$\mathcal{L}_{\text{loc}} = \left\| \hat{\mathbf{M}}_p - \mathbf{M}_p^* \right\|_2^2. \tag{3}$$

The overall Stage-I objective is

$$\mathcal{L}_{\text{PFA}} = \mathcal{L}_{\text{FM}} + w(t)\mathcal{L}_{\text{pix}} + \lambda_{\text{loc}}\mathcal{L}_{\text{loc}}, \tag{4}$$

where $\mathcal{L}_{\text{FM}}$ is the original flow-matching objective of the backbone. By directly supervising the decoded product while jointly learning reference-aware localization, PFA strengthens the model's capability to preserve fine-grained product details.

4.3 Energy-Guided Velocity Correction

Although PFA strengthens the model's general capability to preserve fine-grained product details, this capability is learned across the training distribution and does not explicitly assess fidelity for each reference product during inference. Along an individual sampling trajectory, local product details may still deviate from the reference, while the original denoising process provides no reference-specific feedback to detect and correct such deviations. This mismatch between learned preservation capability and sample-specific fidelity motivates **Energy-Guided Velocity Correction (EVC)**. EVC explicitly evaluates the current generated product against the reference during sampling and uses the resulting fidelity signal to correct the generation trajectory.

At each EVC-enabled sampling step, we obtain the predicted clean latent $\hat{\mathbf{z}}_0$ and its decoded image $\hat{\mathbf{I}}_0$ using Eq. 1. The reference-aware localizer learned in Stage I identifies the current generated-product region. Since the product may undergo changes in scale and placement when integrated into the poster, strict pixel-wise comparison with the reference is unreliable. We therefore measure their discrepancy through local correspondences between the generated product region and the full reference-product image.

Each local position is represented by a detail descriptor combining RGB appearance, edge responses, and high-frequency residuals. Let $\mathbf{q}_i$ and $\mathbf{k}_j$ denote the descriptors at the i-th generated-product location and the j-th reference location, respectively. We denote their normalized squared descriptor distance by d_{ij} and further define

$$\tilde{d}_{ij} = d_{ij} + \lambda_c \|\mathbf{p}_j\|_2^2, \tag{5}$$

where $\mathbf{p}_j \in [-1, 1]^2$ is the normalized coordinate of the j-th reference location. The second term introduces a weak spatial prior that discourages correspondence with peripheral reference regions.

We then define the bidirectional product-fidelity energy as

$$E_{\text{fid}} = \frac{1}{N_g} \sum_i \text{smin}_{\tau,j}\, \tilde{d}_{ij} + \lambda_r \sum_j \omega_j\, \text{smin}_{\tau,i}\, d_{ij}, \tag{6}$$

where smin_τ denotes a differentiable soft minimum and ω_j is the normalized detail-saliency weight of the j-th reference location. The generated-to-reference term suppresses distorted or spurious details, while the reverse term encourages salient reference details that may otherwise be missing from the generated product to remain represented. Detailed definitions of the descriptors, correspondence distance, spatial prior, and reference-detail weighting are provided in Appendix D.2.

Since E_{fid} is differentiable with respect to the predicted clean latent through the frozen VAE decoder, its gradient provides a reference-specific signal for correcting the sampling trajectory:

$$\mathbf{g}_t = \nabla_{\hat{\mathbf{z}}_0} E_{\text{fid}}, \qquad \tilde{\mathbf{v}}_t = \mathbf{v}_t + \frac{\eta}{\max(t, t_{\min})} \bar{\mathbf{g}}_t, \tag{7}$$

where v_t and $\tilde{v}_t$ denote the original and corrected sampling velocities, η controls the correction strength, and $t_{\min}$ stabilizes the update near $t = 0$. $\bar{g}_t$ is the product-region energy gradient after normalization and clipping. The derivation is provided in Appendix D.3.

Table 2: Quantitative comparison on *Fine-grained Fidelity* setting of FidPoster-Bench. The best and second-best results are highlighted in **bold** and underline, respectively. Relative improvements over the strongest competing results are reported where FidPoster ranks first.

		Text Accuracy	Product Fidelity				Poster Quality	
Method	**Venue/Year**	**P-Text↑**	**SSIM↑**	**GMSD↓**	**Color↓**	**HF↓**	**Poster Text↑**	**FID↓**
Qwen-Image-Edit	Tech. Rep. 2025	71.51	0.6682	0.1772	8.66	1.112	84.78	39.02
BAGEL	arXiv 2025	44.81	0.3588	0.2731	22.17	1.466	8.83	47.41
FLUX.2	Tech. Rep. 2025	75.15	0.4733	0.2397	15.57	1.482	35.73	42.08
IC-Custom	ICLR 2026	65.54	0.4440	0.2481	16.32	1.363	10.52	47.85
PosterOmni	CVPR 2026	46.89	0.4340	0.2603	16.46	1.371	86.40	46.55
OmniGen2	CVPR 2026	44.35	0.3869	0.2729	20.59	1.274	9.16	59.12
Step1X-Edit	arXiv 2025	78.22	0.7588	0.1378	8.34	0.981	15.87	61.47
FidPoster	Ours	**85.49**	**0.8630**	**0.1066**	**3.379**	**0.971**	**86.61**	**34.12**
Improvement	–	*9.29%↑*	*13.73%↑*	*22.64%↓*	*59.48%↓*	*1.02%↓*	*0.24%↑*	*12.56%↓*

Table 3: Human evaluation on *Independent Reference* setting of FidPoster-Bench.

Method	**Product Fidelity**	**Layout and Composition**	**Visual Quality**	**Commercial Usability**	**Overall**
BAGEL	2.789	1.569	1.629	1.372	1.840
FLUX.2	4.044	2.550	2.433	1.981	2.752
IC-Custom	3.773	2.259	2.271	1.844	2.537
OmniGen2	2.283	1.307	1.437	1.275	1.576
PosterOmni	3.920	3.785	3.631	3.354	3.673
Qwen-Image-Edit-2509	4.643	4.111	4.009	3.895	4.165
Step1X-Edit	4.461	1.554	1.559	1.450	2.256
FidPoster (Ours)	**4.838**	**4.326**	**4.244**	**4.235**	**4.411**

5 Experiments

5.1 Experimental Setup

Evaluation Protocol. We evaluate the *Fine-grained Fidelity* samples of FidPoster-Bench with automatic metrics using the same GroundingDINO–SAM2 Liu et al. (2024); Ravi et al. (2025) product-extraction pipeline for all methods. Product Text Score measures OCR-based Cui et al. (2025) packaging-text accuracy; SSIM, GMSD Xue et al. (2013), Color, and HF measure aligned product fidelity; Poster Text Score evaluates promotional-copy accuracy; and FID Heusel et al. (2017) evaluates overall poster quality. Metric details are provided in Appendix E.1. On the *Independent Reference* samples, five annotators blindly rate randomized and anonymized results from 1–5 in product fidelity, layout and composition, visual quality, and commercial usability. Each criterion contains 1,000 ratings per method, and Overall averages the four criteria.

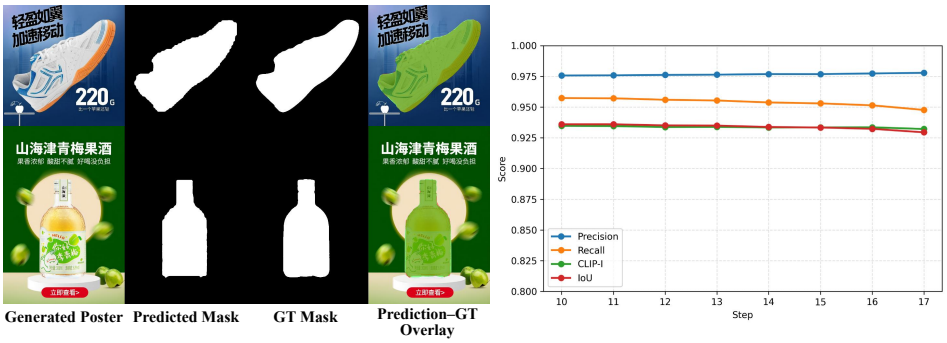
Implementation Details. FidPoster is initialized from Qwen-Image-Edit-2509 Wu et al. (2025) and fine-tuned for 4 epochs with rank-16 LoRA, AdamW8bit, a learning rate of 5×10^{-5}, and BF16 precision. The Charbonnier, Sobel, and MS-SSIM weights are $1.0/0.2/0.2$, with $\gamma = 0.6$ and localization weight 0.1. Inference uses 20 steps and CFG 4.0; EVC is applied over 55%–90% of the trajectory with guidance scale 0.60 and maximum gradient norm 5.0. Edge and high-frequency correspondence weights are both 1.0, and the reference-coverage weight is 0.8. Training uses 8 NVIDIA A40 48GB GPUs for approximately 28 hours.

5.2 Main Results

Automatic Evaluation. Table 2 shows that FidPoster consistently outperforms all baselines in both product fidelity and poster quality. Compared with Step1X-Edit Liu et al. (2025), FidPoster improves Product Text Score and SSIM by 9.29% and 13.73%, while reducing GMSD and color difference by 22.64% and 59.48%. Detection failures remain low across methods, ranging from 0.00–2.00% on Text-Fidelity and 0.19–2.65% on Visual-Fidelity, with detailed coverage reported in Appendix E.2. The higher Product Text Score reflects better preservation of packaging content, while the improvements in SSIM, GMSD, and color difference indicate more accurate structures and



Figure 3: Qualitative comparison on FidPoster-Bench.



(a) Qualitative Localization Results (b) Localization Accuracy and Stability

Figure 4: Evaluation of the reference-aware product localizer. (a) Qualitative comparison between predicted and ground-truth masks. (b) Localization accuracy and stability across sampling steps.

product appearance. The smaller HF gain suggests that both methods retain similar amounts of high-frequency information; however, HF measures frequency statistics rather than whether fine details are semantically correct. The qualitative results in Figure 3 confirm that FidPoster more faithfully reconstructs meaningful details such as characters, logos, textures, and local structures. Moreover, Step1X-Edit Liu et al. (2025) achieves strong product preservation but substantially lower overall poster quality. FidPoster maintains strong performance across both product-fidelity and poster-level metrics, indicating that its fidelity enhancement does not compromise complete poster generation.

Human Evaluation. As shown in Table 3, FidPoster achieves the highest scores across all four criteria, with an overall score of 4.411. Compared with Qwen-Image-Edit Wu et al. (2025), the strongest baseline in human evaluation, FidPoster improves product fidelity from 4.643 to 4.838 and commercial usability from 3.895 to 4.235. Paired sample-level analysis further shows statistically significant improvements across all four criteria, with all 95% confidence intervals of the score

Table 4: Ablation study of PFA on the *Fine-grained Fidelity* samples of FidPoster-Bench.

			Text Fidelity	Visual Fidelity			
Charb.	**Sobel**	**MS-SSIM**	**P-Text↑**	**SSIM↑**	**GMSD↓**	**Color↓**	**HF↓**
–	–	–	82.09	0.7936	0.1353	4.448	1.026
✓	–	–	83.46	0.8182	0.1280	3.738	1.013
✓	✓	–	**85.51**	0.8311	0.1195	3.818	0.977
✓	–	✓	83.93	0.8174	0.1273	3.854	0.977
✓	✓	✓	*84.92*	**0.8561**	**0.1097**	**3.501**	**0.973**

Table 5: Ablation study of EVC on the *Fine-grained Fidelity* samples of FidPoster-Bench.

	Text Fidelity	Visual Fidelity			
Configuration	**P-Text↑**	**SSIM↑**	**GMSD↓**	**Color↓**	**HF↓**
w/o HF	<u>84.74</u>	**0.8632**	<u>0.1067</u>	<u>3.383</u>	0.9830
w/o Ref.	84.70	0.8610	0.1075	3.384	0.9931
w/o Edge	84.37	0.8616	0.1076	3.414	<u>0.9791</u>
Full EVC	**85.49**	<u>0.8630</u>	**0.1066**	**3.379**	**0.9709**

differences above zero. Notably, FidPoster achieves higher overall scores on 70.0% of the evaluated samples. Detailed statistical analysis is provided in Appendix E.3.

Localization and Efficiency. Figure 4 shows stable localization across sampling steps, with Precision near 0.98, Recall around 0.95, and CLIP-I and IoU above 0.92. Appendix E.4 reports only a 6% inference-time overhead over Qwen-Image-Edit, indicating that the additional fidelity guidance remains computationally practical.

5.3 Ablation Studies

Effect of Pixel-Space Supervision. Table 4 shows that Charbonnier supervision improves all metrics over Standard SFT, confirming the value of directly constraining decoded product appearance. Adding Sobel supervision yields the highest Product Text Score, indicating that edge constraints are particularly effective for packaging characters and local boundaries. MS-SSIM provides complementary multi-scale structural supervision. Although the Charbonnier–Sobel combination achieves the best Product Text Score, the full PFA objective obtains the best SSIM, GMSD, Color, and HF results, demonstrating the most balanced preservation of product details.

Effect of Energy-Guided Velocity Correction. As shown in Table 5, full EVC consistently improves the PFA-trained model. Removing the reference cue causes the largest HF degradation, showing its importance for recovering missing reference details, while removing the edge cue most strongly affects Product Text Score and color consistency. Removing the HF cue slightly increases SSIM from 0.8630 to 0.8632 but substantially worsens HF loss, indicating that structural similarity alone does not capture high-frequency fidelity. Full EVC achieves the best results on four of the five metrics and comparable SSIM, supporting the complementarity of the three fidelity cues.

6 Conclusion

We present FidPoster, an end-to-end framework for product poster generation with fine-grained product-detail fidelity. Our analysis shows that VAE compression and reconstruction can degrade fine-grained product details, motivating direct supervision of decoded appearance and reference-guided correction during sampling. FidPoster combines Pixel-Space Fidelity Activation (PFA) for learning decoded product fidelity with Energy-Guided Velocity Correction (EVC) for correcting sample-specific fidelity deviations along the sampling trajectory. We further introduce FidPoster14K for fidelity learning and FidPoster-Bench with complementary Fine-grained Fidelity and Independent Reference settings. Experiments demonstrate consistent improvements in product fidelity across both automatic and human evaluations.

AI USE STATEMENT

Generative AI tools were used to assist with language polishing and improving the clarity of the manuscript. They were also used during dataset construction to assist with automatic data screening and the generation of detailed design instructions from poster images. All automatically processed data and generated annotations were subsequently manually inspected and verified by the authors. The research ideas, methodological design, experimental settings, evaluation procedures, quantitative results, analyses, and scientific conclusions were developed and verified by the authors, who take full responsibility for the final manuscript and dataset.

REFERENCES

Yalong Bai, Yuxiang Chen, Wei Yu, Linfang Wang, and Wei Zhang. Products-10k: A large-scale product recognition dataset. *arXiv preprint arXiv:2008.10545*, 2020.

Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Aiming Hao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Yuming Jiang, et al. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025.

Haoyu Chen, Xiaojie Xu, Wenbo Li, Jingjing Ren, Tian Ye, Songhua Liu, Ying-Cong Chen, Lei Zhu, and Xinchao Wang. Posta: A go-to framework for customized artistic poster generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28694–28704, 2025.

Sixiang Chen, Jianyu Lai, Jialin Gao, Hengyu Shi, Zhongying Liu, Tian Ye, Junfeng Luo, Xiaoming Wei, and Lei Zhu. Posteromni: Generalized artistic poster creation via task distillation and unified reward feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5978–5987, 2026.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Benlei Cui, Fangao Zeng, Weitao Jiang, Yuwen Zhai, Haiwen Hong, Longtao Huang, Hui Xue, Wenxiang Shang, and Pipei Huang. simpleposter: a simple baseline for product poster generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14854–14863, 2026.

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.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Jiahao Fan, Yuxin Qin, Wei Feng, Yanyin Chen, Yaoyu Li, Ao Ma, Yixiu Li, Li Zhuang, Haoyi Bian, Zheng Zhang, et al. Autopp: Towards automated product poster generation and optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 3768–3776, 2026.

Yifan Gao, Jinpeng Lin, Min Zhou, Chuanbin Liu, Hongtao Xie, Tiezheng Ge, and Yuning Jiang. Textpainter: Multimodal text image generation with visual-harmony and text-comprehension for poster design. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 7236–7246, 2023.

Yifan Gao, Zihang Lin, Chuanbin Liu, Min Zhou, Tiezheng Ge, Bo Zheng, and Hongtao Xie. Postermaker: Towards high-quality product poster generation with accurate text rendering. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8083–8093, 2025.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Xiwei Hu, Haokun Chen, Zhongqi Qi, Hui Zhang, Dexiang Hong, Jie Shao, and Xinglong Wu. Dreamposter: A unified framework for image-conditioned generative poster design. *arXiv preprint arXiv:2507.04218*, 2025.

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2, 2025.

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. URL https://arxiv.org/abs/2506.15742.

Zhaochen Li, Fengheng Li, Wei Feng, Honghe Zhu, Yaoyu Li, Zheng Zhang, Jingjing Lv, Junjie Shen, Zhangang Lin, Jingping Shao, et al. Planning and rendering: Towards product poster generation with diffusion models. *arXiv preprint arXiv:2312.08822*, 2023.

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*, pp. 1250–1260, 2023.

Junle Liu, Peirong Zhang, Yuyi Zhang, Pengyu Yan, Hui Zhou, Xinyue Zhou, Fengjun Guo, and Lianwen Jin. Posterverse: A full-workflow framework for commercial-grade poster generation with html-based scalable typography. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 7197–7205, 2026.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024.

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025.

Jian Ma, Yonglin Deng, Chen Chen, Nanyang Du, Haonan Lu, and Zhenyu Yang. Glyphdraw2: Automatic generation of complex glyph posters with diffusion models and large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 5955–5963, 2025.

Yuxin Qin, Ke Cao, Haowei Liu, Ao Ma, Fengheng Li, Honghe Zhu, Zheng Zhang, Run Ling, Wei Feng, Xuanhua He, et al. Innoads-composer: Efficient condition composition for e-commerce poster generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 32988–32999, 2026.

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. In *International Conference on Learning Representations*, volume 2025, pp. 28085–28128, 2025.

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.

Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. *arXiv preprint arXiv:2109.07161*, 2021.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Super Intelligence Team, Changhao Qiao, Chao Hui, Chen Li, Cunzheng Wang, Dejia Song, Jiale Zhang, Jing Li, Qiang Xiang, Runqi Wang, et al. Firered-image-edit-1.0 technical report. *arXiv preprint arXiv:2602.13344*, 2026.

Amogh Vinaykumar, Aiden Li, Suozhi Huang, and Shilong Liu. Any2poster: Any-source poster generation across modalities and domains. *arXiv preprint arXiv:2606.02915*, 2026.

Shaodong Wang, Yunyang Ge, Liuhan Chen, Haiyang Zhou, Qian Wang, Xinhua Cheng, and Li Yuan. Prompt2poster: Automatically artistic chinese poster creation from prompt only. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 10716–10724, 2024.

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.

Wufeng Xue, Lei Zhang, Xuanqin Mou, and Alan C Bovik. Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. *IEEE transactions on image processing*, 23(2): 684–695, 2013.

A Analysis of Product-Detail Distortion

To further analyze the product-detail distortion discussed in the main paper, we examine two representative models with complementary scopes: Qwen-Image-Edit Wu et al. (2025), a general-purpose image editing model, and PosterOmni Chen et al. (2026), a poster-specific generation model. We analyze their failure statistics and representative cases to examine how frequently fine-grained product details are distorted in generated product posters.

A.1 Evaluation Protocol

We randomly sample the same 200 examples from the *Text-Fidelity* subset of **Fine-grained Fidelity** setting in **FidPoster-Bench** for each model and separately assess promotional-text correctness and product-detail correctness. A result is considered *text correct* when the intended promotional copy is correctly rendered in the generated poster. Product correctness is determined by comparing the generated product with the reference product. A result is labeled as *product incorrect* if at least one visually distinctive product detail, such as packaging text, logos, numerical specifications, or local textures, is evidently missing, distorted, or altered.

A.2 Prevalence of Product-Detail Distortion

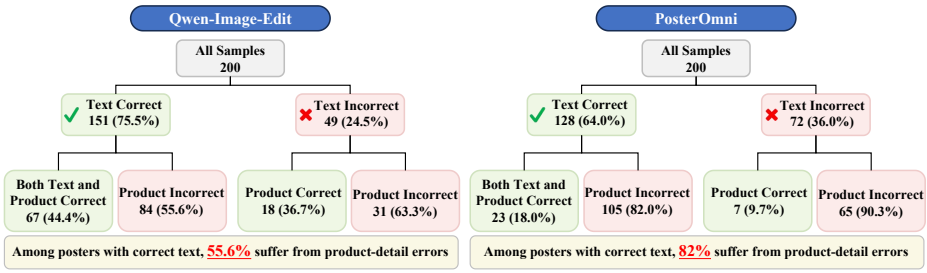


Figure 5: Statistical breakdown of text correctness and product-detail correctness for Qwen-Image-Edit and PosterOmni on 200 evaluated samples. Percentages at each child node are computed with respect to the samples in its corresponding parent node.

Figure 5 presents the statistical breakdown of promotional-text correctness and product-detail correctness for Qwen-Image-Edit Wu et al. (2025) and PosterOmni Chen et al. (2026). For Qwen-Image-Edit Wu et al. (2025), 151 of the 200 generated posters contain correct promotional text. Among these results, 67 (44.4%) also preserve the product details correctly, whereas 84 (**55.6%**) still exhibit product-detail distortions. For PosterOmni Chen et al. (2026), 128 results contain correct promotional text, but only 23 (18.0%) preserve the product correctly, while 105 (**82.0%**) contain product-detail errors.

Product-detail distortion also occurs frequently when considering all generated samples. Overall, 115 of the 200 Qwen-Image-Edit Wu et al. (2025) results (**57.5%**) and 170 of the 200 PosterOmni Chen et al. (2026) results (**85.0%**) contain product-detail errors. Similar patterns are observed in the text-incorrect branches, where **63.3%** of Qwen-Image-Edit results Wu et al. (2025) and **90.3%** of PosterOmni Chen et al. (2026) results also exhibit product-detail distortions. These statistics indicate that fine-grained product-detail distortion occurs at a high frequency in both a general-purpose image-editing model and a poster-specific generation model.

Importantly, many of these failures occur in posters whose promotional content has already been rendered correctly. In such cases, the poster may appear visually complete at first glance, while critical product information, such as packaging text, logos, numerical specifications, or local visual details, is still corrupted. Such errors can substantially reduce the commercial usability of the generated poster, since inaccurate product details may misrepresent the original product even when the overall design remains visually plausible. Together, these results suggest that product-detail fidelity is not an occasional artifact of a particular model, but a recurring and practically important challenge across the two representative generation paradigms examined here.

A.3 Representative Failure Cases



Figure 6: Representative product-detail failure cases produced by Qwen-Image-Edit and PosterOmni. **Green** boxes mark correct textual details in the reference products, **red** boxes indicate distorted text in the generated products, and **blue** boxes indicate textual information missing from the generated products.

Figure 6 presents representative examples from the product-incorrect cases identified in our analysis. The observed errors predominantly occur in fine-grained local regions rather than affecting the overall product identity. Common failure patterns include missing or malformed packaging charac-

ters, corrupted numerical specifications, distorted logos, and alterations of local graphic or texture details. In many cases, the generated product remains semantically recognizable, while these local discrepancies make its appearance inconsistent with the reference product. These examples further illustrate that preserving the overall product identity does not necessarily ensure faithful preservation of its fine-grained details.

B SAMPLES OF FIDPOSTER14K



Figure 7: Representative samples illustrating the structure of **FidPoster14K**. Each sample consists of a target poster, a reference product image, a detailed design instruction, and a product mask.

FidPoster14K is a large-scale dataset for product-conditioned commercial poster generation, containing 14,067 multimodal samples. Each sample consists of a reference product image, its corresponding target poster, a detailed design instruction, and a product mask. The dataset covers a broad range of common commercial products, poster layouts, and visual styles. Representative samples and the sample structure are shown in Figure 7.

C DETAILS OF FIDPOSTER-BENCH

Table 6: Composition of FidPoster-Bench under the two evaluation settings.

Setting / Subset	**# Samples**	**Evaluation Focus**
Fine-grained Fidelity	800	Automatic paired evaluation
Text-Fidelity	400	Packaging-text fidelity
Visual-Fidelity	529	Visual-detail fidelity
Independent Reference	200	Poster generation capability
Overall	**1,000**	–

FidPoster-Bench contains 1,000 samples under two complementary evaluation settings: *Fine-grained Fidelity* and *Independent Reference*. The former provides controlled reference-target correspondence for fine-grained automatic evaluation, while the latter evaluates whether fidelity preservation generalizes to independently sourced product images.

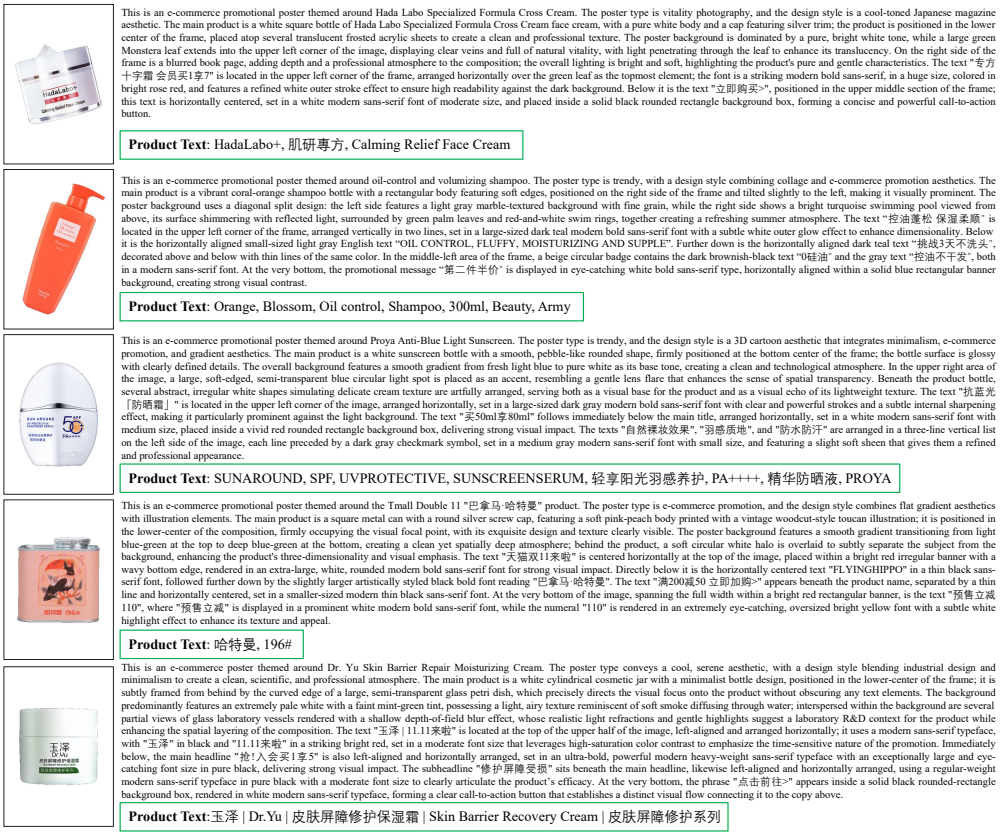


Figure 8: Representative samples from the *Text-Fidelity* subset of **Fine-grained Fidelity Setting**.

C.1 Fine-grained Fidelity Setting

A randomly sampled poster set may contain many products with limited fine-grained visual information, making it insufficient for systematically evaluating product-detail fidelity. We therefore construct the *Fine-grained Fidelity* setting by manually selecting and verifying 800 held-out product–poster pairs whose reference products contain clearly identifiable fine-grained details. These samples are fully disjoint from FidPoster14K and provide paired target posters for controlled evaluation of whether product-specific details are faithfully preserved in generated posters.

During benchmark selection, we inspect the reference product of each candidate sample and retain samples containing visually identifiable details that can be reliably compared with the generated product. We exclude cases in which the reference product is severely blurred, heavily occluded, or lacks sufficiently distinctive local information for fidelity evaluation. Each retained sample is manually verified to ensure the validity of the reference product and its correspondence with the target poster.

According to the types of fine-grained details present in the reference products, the selected samples are further organized into two partially overlapping subsets: *Text-Fidelity* and *Visual-Fidelity*.

Text-Fidelity. The *Text-Fidelity* subset contains 400 samples whose reference products contain clearly visible textual information, such as product names, packaging copy, numerical specifications, or other printed text. These samples are used to evaluate whether fine-grained textual details on the reference product are faithfully preserved in the generated poster.

Visual-Fidelity. The *Visual-Fidelity* subset contains 529 samples with distinctive non-textual product details, including surface textures, colors, local patterns, and structural characteristics. This subset evaluates the preservation of fine-grained visual appearance beyond textual information.



This is an e-commerce poster themed around the "怪兽小新" cartoon phone case, with a trendy poster type and a 2D animation design style that combines impasto and flat design aesthetics, featuring a clean and powerful composition. The main product is a phone case with a cartoon illustration style, its surface exhibiting an impasto oil painting-like brushstroke texture, placed vertically in the left area of the frame, with its wavy edges and camera cutout details clearly visible. The poster background adopts a bold vertical color block segmentation design; the left side is deep cyan-blue, providing a steady backdrop for the product, while the right side is bright cyan-blue, offering an open and pure display space for textual information. The overall lighting is even, creating a modern and crisp graphic visual feel. The text "新款," and "怪兽小新" are located in the upper right portion of the frame, arranged vertically in two lines; the font is a modern sans-serif bold black typeface, extra-large in size and highly eye-catching; the color is pure white without any stroke effects, forming a strong contrast against the cyan-blue background for significant visual impact, and there is no background box. The text " 经典黑" is positioned directly below the main title, arranged horizontally; the font is the same modern sans-serif bold black typeface as the main title, large in size but slightly smaller than the main title; the color is dark gray, distinguishing information hierarchy through color variation, also without a background box. The text "防滑波浪边" is located in the lower right portion of the frame, arranged horizontally; the font is a medium-weight sans-serif typeface, moderate in size; the color is pure white, placed inside a light gray rounded rectangular capsule-shaped background box with a subtle three-dimensional effect, making it appear to float slightly above the background.



This is a size comparison e-commerce poster themed around high-quality beige casual men's shoes. The poster type is minimalist e-commerce display, and the design style integrates minimalism, old money aesthetic, and still life commercial photography aesthetics, emphasizing the purity of materials and composition. The core subject of the image is a beige fabric casual shoe with excellent texture, firmly positioned in the lower-middle section vertically and centered horizontally within the frame. It is captured from a standard side profile view with the toe pointing toward the left side of the image. The canvas texture of the upper, tone-on-tone laces, and rubber texture of the sole are clearly visible under soft lighting, achieving an extremely high level of realism. The background design adopts a minimalist negative space strategy, simulating the seamless wall effect of a high-end photography studio. The entire background is covered in a delicate light warm gray tone with highly natural light and shadow transitions. Only a soft contact shadow is retained beneath the sole to convey volume, while the large area of negative space above guides the viewer's focus toward the product. The visual center of gravity is lowered to the bottom, where a dark coffee-colored rectangular horizontal bar spanning the left and right edges of the frame serves as a base, both dividing the space and carrying key information. The text "本款40码 = 运动鞋41码 = 皮鞋40码" is precisely placed at the vertical and horizontal center of the bottom dark brown color block. The font selected is a clean, crisp white modern sans-serif bold typeface with uniform and powerful stroke weight, free of any unnecessary decoration or effects, utilizing extremely high color contrast to achieve optimal readability. The overall image achieves perfect balance through asymmetrical negative space and a solid color block at the bottom, conveying a comfortable, professional, and trustworthy brand image.



This is an e-commerce poster themed around the high-end Seagull brand mechanical women's watch. The poster category belongs to popular Guofeng (Chinese style), with an overall design style that blends vitality photography and Atmospheric Vibe, resulting in a gorgeous and elegant visual aesthetic. The absolute visual center of the composition is a vertically placed skeleton mechanical wristwatch, positioned directly in the center of the frame. The case is crafted from noble rose gold, with flowing surface luster, paired with a dark brown leather strap featuring delicate textures. The intricate golden gear mechanism inside the dial is clearly visible, precisely interlocked, showcasing the beauty of industrial art. The background is designed as an immersive deep red tone with rich layering, surrounded by several blooming red peonies with layered, complex petals and intense colors. The flowers on the right side and edges are blurred, creating a strong contrast between virtual and real elements against the sharp watch subject, evoking a sense of deep space and romantic atmosphere without obscuring any text information. The text "浮光掠影 曳彩动人" is located at the top center of the image, arranged horizontally, using an elegant white Songti font with thin, sharp strokes and extended strokes, in a huge size with high recognizability, forming a sharp contrast against the red background. The text "「大女主」镂空机械女表" is placed closely below the main title, centered horizontally, using a white modern thin sans-serif Heiti font, with moderate size, concise and clean. The entire poster perfectly interprets the "Big Heroine" temperament of modern women, combining independence and confidence with softness, through the strong contrast of the red and gold color scheme.



This is an e-commerce poster themed around a silent mouse. The poster type is digital product, and the design style is minimalism, 3D product rendering, soft tones, and e-commerce promotion. The main product is a pink-and-white esports-style mouse with an angular, modern exterior design, positioned in the lower-center-right area of the frame, displayed from a front-side angle. Several rings of white semi-transparent concentric circular ripple graphics are overlaid above the left and right click buttons of the mouse, symbolizing the silent clicking effect, and a clear reflection is cast beneath the product to enhance its three-dimensionality. The overall background of the poster presents a uniform, soft dusty pink matte texture, with a partial view of a white laptop—including the keyboard and trackpad areas—extending into the top-right edge of the frame; this element is heavily blurred as an environmental backdrop, creating a space that is both clean and imbued with a sense of everyday life, illuminated by bright, soft lighting without obvious shadows. A semi-transparent bright cyan-blue circular geometric color block decorates the top-left corner of the image. The text "低音按键" is located in the top-left corner of the frame, adjacent to the blue dot, arranged horizontally in a pure white modern bold sans-serif font with a very large and eye-catching size. The text "不打扰他人" is positioned directly below the preceding headline, arranged horizontally in a pure white modern bold sans-serif font, also in a very large size. The text "按键采用低音技术" is placed below the main headline, aligned horizontally to the left in a white modern medium-weight sans-serif font with a moderate size. The text "不打扰舍友, 朋友, 亲人休息" is situated below the preceding line of text, arranged horizontally in a white modern medium-weight sans-serif font. The text "享受健康生活" is located at the bottom of this group of copy, arranged horizontally in a white modern medium-weight sans-serif font; all text elements are clearly distributed across the upper-left portion of the pink background, forming a perfect diagonal balance with the product in the lower-right area.



This is an e-commerce poster themed around a pre-sale promotion for children's cartoon storage cabinets. The poster type is e-commerce promotional, and the design style combines vibrant photography with 3D cartoon aesthetics. The main products are two teal-colored, cartoon-style children's storage cabinets placed side by side, with the taller unit on the left and the shorter one on the right, both positioned in the lower-center area of the composition; the cabinet drawers feature an adorable white duckling design, and the open compartments above are filled with children's books. The background depicts a bright and cozy corner of a children's room, with a large white window directly behind the main subjects allowing soft light to filter through, creating an airy and clean atmosphere, while the view outside the window is rendered as pure white; on the left side of the frame, a corner of a bean-paste-colored curtain is visible, and the floor is covered with an off-white rug featuring a colorful speckled pattern, all captured with a shallow depth-of-field effect to keep focus on the products. The text "预售付定金 放心购" is centered in the upper half of the image, arranged horizontally in a very large, true-red, modern bold sans-serif font without any special effects, overlaid directly onto the bright background. The text "至高立省12.7元" is positioned directly below the main headline, arranged horizontally within a red rounded-rectangle label in white sans-serif typography, with the numerals "12.7" specially enlarged and bolded to emphasize the discount; the label features subtle three-dimensional highlights to enhance visual hierarchy.



This is an e-commerce poster themed around a promotional campaign for a purple hair dryer, with the poster type being electronic Chinese style. The design style integrates 3D cartoon aesthetics, minimalism, and industrial design, presenting an overall cool and premium feel. The main product is a T-shaped purple hair dryer featuring a delicate matte frosted texture on the body, while the joint between the handle and the barrel showcases a smooth, reflective electroplated metallic purple finish; it is equipped with a black comb nozzle at the bottom. The product stands vertically in the center-left of the frame, appearing sleek and highly technological. The background constructs a bright, pure three-dimensional space where an extremely light lavender-gray wall on the left meets a pale purple wall and floor on the right to form a corner, creating a soft sense of depth. A cluster of pale purple campanula flowers extends into the mid-ground from the right side, adding natural vitality and elegance. In the foreground, three crystal-clear glass bubble spheres are scattered across the floor, reflecting bright luster that echoes the theme of "luster." The text "以风养发五部曲" is positioned horizontally at the top center of the image, rendered in a massive, highly conspicuous modern sans-serif bold font colored in a steady deep indigo-purple, featuring an extremely subtle soft white inner glow effect to enhance its three-dimensionality. The text "让秀发焕发光泽" is placed directly below the main headline, arranged horizontally in a medium-sized modern sans-serif font matching the color of the main title, with crisp and sharp edges emphasizing a premium minimalist style through its pure texture.

Figure 9: Representative samples from the *Visual-Fidelity* subset of **Fine-grained Fidelity Setting**.

The two subsets are not mutually exclusive, since real-world products often contain both textual and non-textual fine-grained details. This overlapping design allows different fidelity dimensions to be evaluated without artificially assigning each product to a single category. Poster Text Score and FID are additionally evaluated on all 800 Fine-grained Fidelity samples to assess promotional-copy accuracy and overall poster quality.

C.2 Independent Reference Setting

Although the Fine-grained Fidelity setting provides controlled reference–target correspondence for fine-grained fidelity analysis, real-world poster generation typically starts from an independently provided product image rather than a product extracted from a paired target poster. Evaluation restricted to paired samples therefore cannot fully reflect whether the learned fidelity-preserving capability generalizes to such practical inputs.

To complement the Fine-grained Fidelity setting, we construct an *Independent Reference* setting containing 200 product images independently sourced from Products-10K Bai et al. (2020). Each product is paired with a design instruction and directly used as the reference input for poster generation. These products are independent of the poster sources used to construct FidPoster14K and the Fine-grained Fidelity benchmark, providing a complementary setting for examining fidelity preservation beyond controlled reference–target pairs.



This is an e-commerce promotional poster themed around a moisturizing body lotion. The poster type is fresh minimalism skincare promotion, and the design style combines clean commercial product photography with a soft botanical aesthetic, emphasizing purity, hydration, and gentle care. The main subject is a white cylindrical pump bottle of MINGOU body lotion, positioned slightly to the right of the center in the frame. The bottle features a sleek black pump dispenser and a label adorned with delicate golden line-art leaf patterns, clearly displaying the brand name 'MINGOU' and the product name 'MOISTURE BODY LOTION'. The bottle stands on a pristine white surface with a subtle, soft reflection beneath it, enhancing the sense of premium quality. The background is a smooth, airy gradient transitioning from pure white at the bottom to a very pale, refreshing mint-cream green at the top, creating a breathable and hydrating atmosphere. To the left of the bottle, occupying the negative space, is the main copy area. A few translucent, soft-focus green leaves float gently in the upper left and lower right corners, adding depth and reinforcing the natural ingredients concept without obscuring the product. The text \"深层滋润\" is located in the upper left quadrant, arranged horizontally in a large, elegant dark forest green modern sans-serif bold font, serving as the primary headline. Directly below it, the text \"雪肌倍滑\" follows in the same dark green font and size, aligned left. Beneath these headlines, the text \"保湿滋润 · 防裂止痒\" is placed inside a soft golden-yellow capsule-shaped badge with rounded ends, using a smaller dark brown sans-serif font to highlight key benefits. Further down on the left, the text \"持久留香 清爽不油腻\" is arranged horizontally in a medium-sized grey-green modern sans-serif font, providing additional product details. The overall layout balances the visual weight between the text on the left and the product on the right, utilizing a fresh color palette of white, pale green, and gold to convey a sense of natural luxury and effective skincare.



This is an e-commerce promotional poster themed around classic American vintage brown leather work boots. The poster type is high-end product display, and the design style combines rugged workwear aesthetics with clean, minimalist commercial photography, emphasizing material texture and craftsmanship. The main subject is a single reddish-brown crazy horse leather boot, positioned prominently in the lower-right center of the frame, angled at a three-quarter view with the toe pointing towards the bottom left. The boot features distinct yellow and brown woven laces, silver metal eyelets, visible white contrast stitching along the moc-toe and welt, and a thick cream-colored wedge sole, showcasing a worn-in, premium leather texture under soft, directional studio lighting that highlights the natural grain and subtle scuffs of the hide. The background is a textured, warm beige surface resembling heavy kraft paper or a raw concrete wall, providing a neutral yet atmospheric backdrop that complements the earth tones of the product without distraction. A soft, natural shadow is cast beneath the sole to ground the object. The text \"复古工装\" is located in the upper left quadrant, arranged horizontally in a large, bold, dark charcoal sans-serif font, serving as the main headline. Directly below it, the text \"经典楔形底\" appears in a slightly smaller but equally bold dark charcoal font, aligned left. Further down on the left side, a vertical arrangement of selling points includes the text \"头层疯马皮\" enclosed in a thin dark rectangular outline, and below that, the text \"耐磨防滑\" in a medium-sized clean sans-serif font. The overall composition utilizes negative space on the left for typography while anchoring the visual weight with the detailed product on the right, creating a balanced, professional, and masculine aesthetic suitable for fashion retail.



This is an e-commerce promotional poster themed around a children's antibacterial foaming hand wash. The poster type is fresh maternal and infant care promotion, and the design style combines minimalist product photography with a playful, cute cartoon aesthetic, emphasizing cleanliness, safety, and child-friendliness. The main subject is a white pump-dispenser bottle of hand wash with a bright blue pump head and collar, positioned prominently in the center-right of the frame. The bottle features a gradient pink-to-white label adorned with cute cartoon illustrations of a white rabbit and a green parrot, along with visible bubble graphics and product text. The bottle appears three-dimensional and pristine, resting on a subtle reflective surface that suggests purity. The background features a soft, dreamy gradient transitioning from a clean sky blue at the top to a gentle pastel pink at the bottom, evoking a sense of soft foam and gentle care. Several translucent, iridescent soap bubbles of varying sizes float dynamically around the bottle, particularly in the upper left and lower right areas, adding depth and reinforcing the 'foaming' product attribute. The lighting is bright, soft, and diffused, typical of high-end studio product photography, creating specular highlights on the plastic bottle and bubbles to enhance texture. The text \"长效保护\" is located in the upper left quadrant, arranged horizontally in a large, rounded, bold white sans-serif font with a soft blue drop shadow to ensure readability against the light background. Directly below it, the text \"温和配方\" appears in a slightly smaller, medium-weight dark blue sans-serif font, left-aligned with the headline. Further down on the left side, the text \"绵密泡沫\" is enclosed within a white capsule-shaped badge with a thin blue outline, using a blue bold sans-serif font. In the bottom left corner, the text \"呵护稚嫩双手\" is arranged horizontally in a small, clean dark gray sans-serif font, serving as a supportive tagline. The overall layout balances the product on the right with the typographic hierarchy on the left, utilizing negative space effectively to create a breathable, hygienic, and appealing visual for parents.



This is an e-commerce promotional poster themed around a camouflage off-road toy jeep. The poster type is dynamic product display, and the design style combines rugged outdoor aesthetics with clean commercial photography, emphasizing power and playfulness. The main subject is a green and yellow camouflage patterned toy monster truck with large black knobby tires, olive-green rims, a detailed black front bull bar, and a roof rack with spotlights. The vehicle is positioned in the lower-right center of the frame, angled slightly towards the left in a three-quarter view to showcase its aggressive stance and tread details. It appears to be driving over a subtle, stylized rocky terrain texture at the very bottom, kicking up a faint, stylized dust effect to imply motion. The background features a gradient from deep forest green at the top to a lighter, misty sage green at the bottom, creating an atmospheric jungle or野外 environment without distracting from the product. Soft, directional studio lighting highlights the glossy finish of the car body and the matte texture of the tires. The text \"越野先锋\" is located in the upper left corner, arranged horizontally in a heavy, impact-style white font with a slight black drop shadow for readability against the green background. Below it, the text \"征服全地形\" is aligned left in a bold, bright yellow sans-serif font, creating a strong visual pop. Further down on the left side, three small icon-text combinations are arranged vertically: a tire icon with \"防滑轮胎\", a spring icon with \"强力减震\", and a shield icon with \"耐撞车身\", all in white modern sans-serif text. The overall layout balances the heavy visual weight of the car on the right with the bold typography on the left, conveying excitement and durability.

Figure 10: Representative samples from the **Independent Reference Setting**.

C.3 Representative Samples

Figure 8 presents representative samples from the *Text-Fidelity* subset of the *Fine-grained Fidelity* setting. These products contain diverse textual details, including brand names, product names, numerical specifications, capacity information, and packaging copy. For each sample, we annotate the product text used for fine-grained fidelity evaluation. The examples cover variations in language, font size, orientation, and spatial placement, making accurate preservation particularly challenging.

Figure 9 shows representative samples from the *Visual-Fidelity* subset of the *Fine-grained Fidelity* setting. These products contain diverse non-textual characteristics, including shape, color, texture, local patterns, and structural details, providing challenging cases for fine-grained visual fidelity evaluation.

Figure 10 presents representative samples from the *Independent Reference* setting. Unlike the Fine-grained Fidelity samples, these reference products are independently sourced and are not extracted from paired target posters.

D Additional Method Details

D.1 Reference-Aware Product Localization

We provide additional details of the reference-aware product localizer introduced in Section 4.2. Its purpose is to localize the product in the poster stream by measuring its correspondence with the reference product, thereby providing the spatial region required by EVC without introducing an external product mask at inference time. The localization branch operates only on intermediate visual features and leaves the original backbone features unchanged.

Given the intermediate visual features, we first separate the visual tokens according to their corresponding image streams into poster-stream tokens $\mathbf{H}_g$ and reference-product tokens $\mathbf{H}_r$. The two streams are projected into a shared multi-head matching space, where the poster-stream representations serve as queries and the reference-product representations serve as keys. For the i-th poster token and the j-th reference-product token, their similarity on the h-th matching head is computed

as

$$s_{hij} = \frac{\mathbf{q}_{hi}^{\top}\mathbf{k}_{hj}}{\sqrt{d}}, \tag{8}$$

where d denotes the dimension of each matching head. This cross-image matching allows the localization response to depend explicitly on the correspondence between the current poster representation and the reference product, rather than being estimated from the poster stream alone.

A product region in the poster generally corresponds strongly to only a subset of reference-product tokens. We therefore aggregate the matching responses along the reference dimension using log-sum-exp, followed by averaging across matching heads:

$$a_i = \frac{1}{H}\sum_{h=1}^{H}\log\sum_{j=1}^{N_r}\exp(s_{hij}), \tag{9}$$

where H is the number of matching heads and N_r is the number of reference-product tokens. Compared with directly averaging all pairwise responses, this aggregation emphasizes strong local correspondences while remaining differentiable.

The aggregated responses are first normalized independently within each sample:

$$\bar{a}_i = \frac{a_i - \min_k a_k}{\max_k a_k - \min_k a_k + \epsilon}. \tag{10}$$

The normalized responses are then converted into soft localization scores:

$$\hat{M}_{p,i} = \sigma(\beta(2\bar{a}_i - 1)), \tag{11}$$

where $\sigma(\cdot)$ denotes the sigmoid function and β controls the sharpness of the localization response. The resulting scores are reshaped to the spatial poster-token grid to obtain the soft product-location map $\hat{\mathbf{M}}_p$. Consequently, regions exhibiting stronger correspondence with the reference product receive higher localization responses, whereas unrelated poster regions are suppressed.

During training, the ground-truth product mask is bilinearly resized to the corresponding poster-token resolution and processed using the same normalization and soft mapping to obtain the target map $\mathbf{M}_p^*$. The predicted map is supervised using the localization objective defined in Eq. 3. This supervision encourages the reference-aware matching responses to align with the actual product region.

D.2 Product-Fidelity Correspondence Energy

We provide additional details of the product-fidelity correspondence energy introduced in Section 4.3. Its purpose is to measure fine-grained discrepancies between the currently generated product and the reference product without requiring strict pixel-wise alignment.

At each EVC-enabled sampling step, the predicted clean latent is differentiably decoded into image space. The soft product map produced by the reference-aware localizer is upsampled to the image resolution and used to extract the current generated product region. Pixels outside the predicted product region are suppressed, and the resulting crop is resized while preserving its aspect ratio. In contrast, we retain the full reference product image instead of introducing an additional foreground segmentation procedure. The fidelity discrepancy is then measured through local correspondences between the generated product region and the reference image.

To characterize visually distinctive product details, we construct a local descriptor that combines complementary appearance and structural cues. For an image $\mathbf{I}$, the descriptor is defined as

$$\phi(\mathbf{I}) = [\mathbf{I},\ \lambda_e\mathcal{E}(\mathbf{I}),\ \lambda_h\mathcal{H}(\mathbf{I})], \tag{12}$$

where $\mathcal{E}(\cdot)$ extracts local edge responses and $\mathcal{H}(\cdot)$ denotes a high-frequency residual:

$$\mathcal{H}(\mathbf{I}) = \mathbf{I} - \mathrm{AvgPool}_{5\times5}(\mathbf{I}). \tag{13}$$

In practice, $\mathcal{E}(\cdot)$ combines horizontal and vertical Sobel responses with a Laplacian response. The RGB component preserves local color and appearance, while the edge and high-frequency components emphasize packaging text, boundaries, logos, and fine surface structures.

After resizing the descriptor maps to a common correspondence resolution, we flatten them into generated-product descriptors $\{\mathbf{q}_i\}_{i=1}^{N_g}$ and reference descriptors $\{\mathbf{k}_j\}_{j=1}^{N_r}$. Their local discrepancy is measured using the normalized squared Euclidean distance

$$d_{ij} = \frac{\|\mathbf{q}_i - \mathbf{k}_j\|_2^2}{D_\phi}, \tag{14}$$

where D_ϕ denotes the descriptor dimension.

Since the generated product and reference image are not spatially aligned, we do not compare descriptors at identical coordinates. Instead, each local descriptor is matched to its most compatible counterpart through a differentiable soft minimum:

$$\mathrm{smin}_\tau\left(\{d_j\}_{j=1}^{N}\right) = -\tau \log\left(\frac{1}{N}\sum_{j=1}^{N}\exp\left(-\frac{d_j}{\tau}\right)\right), \tag{15}$$

where τ controls the softness of the correspondence.

Because the full reference image may contain peripheral background regions, we introduce a weak spatial prior in the generated-to-reference direction:

$$\tilde{d}_{ij} = d_{ij} + \lambda_c \|\mathbf{p}_j\|_2^2, \tag{16}$$

where $\mathbf{p}_j = (x_j, y_j) \in [-1, 1]^2$ denotes the normalized spatial coordinate of the j-th reference location. This term mildly discourages generated details from being matched to peripheral regions of the reference image. We set $\lambda_c = 0.05$ in our implementation.

A one-way generated-to-reference correspondence encourages every generated local detail to be supported by some region of the reference image, but it does not explicitly penalize important reference details that are absent from the generated product. We therefore define the bidirectional fidelity energy as

$$\mathcal{E}_{\mathrm{fid}} = \frac{1}{N_g}\sum_{i=1}^{N_g}\mathrm{smin}_{\tau,j}\,\tilde{d}_{ij} + \lambda_r \sum_{j=1}^{N_r}\omega_j\,\mathrm{smin}_{\tau,i}\,d_{ij}, \tag{17}$$

where ω_j denotes the normalized detail-saliency weight of reference location j. The first term suppresses distorted or spurious generated structures by requiring them to match the reference, while the reverse term encourages salient reference details to remain represented in the generated product.

The reference saliency weights are derived from local edge and high-frequency responses, with near-white background regions suppressed. Consequently, detail-rich regions such as packaging text, logos, boundaries, and local textures receive greater emphasis in the reference-to-generated coverage term. The reference branch is detached, while $\mathcal{E}_{\mathrm{fid}}$ remains differentiable with respect to the generated clean latent through the frozen VAE decoder. Its gradient therefore provides the sample-specific correction signal used by EVC.

D.3 Derivation and Descent Property of Energy-Guided Velocity Correction

We provide a mathematical derivation of the velocity correction introduced in Section 4.3 and show that it corresponds to a descent direction of the product-fidelity energy in the predicted clean-latent space.

Under the linear flow-matching parameterization, the clean latent associated with the current state $\mathbf{z}_t$ and predicted velocity $\mathbf{v}_t$ is estimated as

$$\hat{\mathbf{z}}_0 = \mathbf{z}_t - t\mathbf{v}_t. \tag{18}$$

Let

$$\mathbf{g}_t = \nabla_{\hat{\mathbf{z}}_0} E_{\mathrm{fid}}(\hat{\mathbf{z}}_0) \tag{19}$$

denote the gradient of the product-fidelity energy defined in Appendix D.2. To reduce this energy, the desired update of the clean prediction is along its negative-gradient direction:

$$\hat{\mathbf{z}}_0' = \hat{\mathbf{z}}_0 - \eta\bar{\mathbf{g}}_t, \tag{20}$$

where $\bar{\mathbf{g}}_t$ denotes the product-region restricted and stabilized gradient used by EVC.

We now seek a corrected velocity $\tilde{\mathbf{v}}_t$ whose corresponding clean prediction realizes Eq. 20. Let $\tilde{\mathbf{v}}_t = \mathbf{v}_t + \Delta\mathbf{v}_t$. Substituting it into Eq. 18 gives

$$\begin{aligned}\tilde{\mathbf{z}}_0 &= \mathbf{z}_t - t\tilde{\mathbf{v}}_t \\ &= \mathbf{z}_t - t(\mathbf{v}_t + \Delta\mathbf{v}_t) \\ &= \hat{\mathbf{z}}_0 - t\Delta\mathbf{v}_t.\end{aligned} \tag{21}$$

Matching Eq. 21 with the desired clean-latent update in Eq. 20 yields

$$\Delta\mathbf{v}_t = \frac{\eta}{t}\bar{\mathbf{g}}_t. \tag{22}$$

Therefore, the corresponding velocity correction is

$$\tilde{\mathbf{v}}_t = \mathbf{v}_t + \frac{\eta}{t}\bar{\mathbf{g}}_t. \tag{23}$$

Importantly, the positive sign in velocity space is not heuristic: because the clean prediction depends negatively on the velocity through $\hat{\mathbf{z}}_0 = \mathbf{z}_t - t\mathbf{v}_t$, adding the energy gradient to the velocity produces a negative-gradient update in clean-latent space.

In practice, directly dividing by t may amplify the correction when t approaches zero. We therefore use the stabilized update

$$\tilde{\mathbf{v}}_t = \mathbf{v}_t + \frac{\eta}{\max(t, t_{\min})}\bar{\mathbf{g}}_t. \tag{24}$$

Substituting Eq. 24 back into the clean-latent estimate gives

$$\tilde{\mathbf{z}}_0 = \hat{\mathbf{z}}_0 - \eta\kappa_t\bar{\mathbf{g}}_t, \qquad \kappa_t = \frac{t}{\max(t, t_{\min})} \in [0, 1]. \tag{25}$$

Hence, $t_{\min}$ only reduces the effective step size near the terminal region and does not change the descent direction.

Descent property. We next show that the practical EVC update remains a descent direction of E_{fid}. Let $\mathbf{M}_t \in [0, 1]$ denote the predicted product mask resized to the clean-latent resolution. The gradient used by EVC is first restricted to the product region and then RMS-normalized and norm-clipped. Since both normalization and clipping only introduce positive scalar rescaling, the processed gradient can be written as

$$\bar{\mathbf{g}}_t = c_t\left(\mathbf{M}_t \odot \mathbf{g}_t\right), \qquad c_t > 0. \tag{26}$$

Using a first-order Taylor expansion of E_{fid} around $\hat{\mathbf{z}}_0$ and Eq. 25, we obtain

$$E_{\text{fid}}(\tilde{\mathbf{z}}_0) = E_{\text{fid}}(\hat{\mathbf{z}}_0) - \eta\kappa_t\mathbf{g}_t^\top\bar{\mathbf{g}}_t + \mathcal{O}(\eta^2). \tag{27}$$

Substituting Eq. 26 gives

$$\mathbf{g}_t^\top\bar{\mathbf{g}}_t = c_t\sum_i M_{t,i}g_{t,i}^2 \geq 0. \tag{28}$$

Therefore,

$$E_{\text{fid}}(\tilde{\mathbf{z}}_0) = E_{\text{fid}}(\hat{\mathbf{z}}_0) - \eta\kappa_t c_t\sum_i M_{t,i}g_{t,i}^2 + \mathcal{O}(\eta^2). \tag{29}$$

For a sufficiently small correction step, the first-order term is non-positive and is strictly negative whenever a non-zero product-region gradient exists. Therefore, under a first-order local approximation, EVC defines a descent direction of E_{fid} in the predicted clean-latent space.

E Additional Experiments

E.1 Detailed Evaluation Protocol for Fine-grained Fidelity Setting of FidPoster-Bench

We provide additional details of the automatic evaluation protocol used for the *Fine-grained Fidelity* setting of FidPoster-Bench. All compared methods are evaluated using the same post-processing and metric computation procedures, with all evaluation parameters fixed across methods. The *Independent Reference* setting is evaluated separately through the human study described in Section 5.1.

Product localization and extraction. For product-region-dependent metrics, we localize the generated product using GroundingDINO followed by SAM2 segmentation. GroundingDINO is queried with the product name associated with each benchmark sample, with both the box and text confidence thresholds set to 0.3. All detected boxes are passed to SAM2, and the resulting masks are merged into a single foreground mask. The merged mask is further refined using morphological closing and hole filling. We obtain the tight foreground bounding box and enlarge it by 10 pixels on each side to retain sufficient local context. The raw bounding-box crop is used for Product Text Score, while the foreground mask is used for visual-fidelity evaluation. Reference-product masks provided by the benchmark are used for the corresponding reference regions. If no valid generated-product region can be extracted, the sample is recorded as a detection failure and excluded only from the affected product-region metric.

Product Text Score. Product Text Score is evaluated on the Text-Fidelity subset using manually verified packaging strings associated with each reference product. We apply PP-OCRv5, using `PP-OCRv5_server_det` and `PP-OCRv5_server_rec`, to the raw bounding-box crop of the generated product. The recognition confidence threshold is set to 0.5, while document orientation classification, document unwarping, and text-line orientation correction are disabled. Before matching, both reference and recognized strings are converted to lower case and normalized by removing spaces, punctuation, and other non-alphanumeric symbols, while retaining numerical and Chinese characters. String similarity is computed using `SequenceMatcher`. To account for fragmented OCR detections, we additionally consider concatenations of at most two recognized text segments, with a penalty of 0.05 applied to each merged pair. Candidate matches with similarity below 0.3 are discarded. We then perform greedy one-to-one matching between recognized and reference strings in descending order of similarity. The final Product Text Score is defined as the accumulated matching similarity normalized by the number of annotated reference strings, such that unmatched reference strings contribute zero to the score.

Product canonicalization and global alignment. For visual-fidelity evaluation, the reference product and the extracted generated product are first independently canonicalized according to their foreground masks. Specifically, we crop each product around its tight foreground bounding box, enlarge the bounding box by 4%, isotropically resize the crop while preserving its aspect ratio, and center-pad it to a 768×768 canonical canvas. We then estimate a single global partial-affine transformation from the generated product to the reference product. SIFT features are extracted within the foreground masks and matched using a ratio threshold of 0.75. When at least six reliable correspondences are available, a partial affine transformation is estimated with RANSAC using a reprojection threshold of 5 pixels, a maximum of 2,000 iterations, and a confidence of 0.99. The estimated transformation is accepted when at least four inliers are obtained. If reliable SIFT correspondences are unavailable, we fall back to a translation that aligns the centers of the two foreground masks. Importantly, the transformation is estimated only once at the whole-product level and is not optimized with respect to any reported evaluation metric. No local or patch-wise registration is performed.

Detail-aware region selection. SSIM, GMSD, and color difference are evaluated on high-detail regions selected exclusively from the reference product. After product canonicalization, we compute a reference-side detail map by equally combining normalized Sobel gradient magnitude and absolute Laplacian response:

$$S_{\text{detail}} = \frac{1}{2}S_{\text{grad}} + \frac{1}{2}S_{\text{lap}}. \tag{30}$$

Candidate regions are sampled as 96×96 patches with a stride of 48. To suppress boundary and background contamination, patch candidates are restricted to an eroded reference-product mask and are retained only when at least 90% of the patch lies inside the product region. Candidate patches are ranked according to their mean detail response, and at most 32 non-redundant regions are retained using center-based suppression. The selected reference patch coordinates are subsequently reused directly on the globally aligned generated product. Therefore, no per-patch nearest-neighbor search or best-similarity matching is performed, preventing misplaced text, logos, or textures from being hidden by matching them to nearby but incorrect regions.

Local structural and appearance metrics. For each selected high-detail region, we compute SSIM, GMSD, and CIEDE2000 color difference at the corresponding coordinates of the aligned

generated product. SSIM and GMSD are evaluated within the reference-product foreground mask, while color difference is computed in CIELAB space using CIEDE2000. The patch-level measurements are aggregated using the reference-side detail responses as weights. Consequently, the reported SSIM, GMSD, and Color metrics place greater emphasis on visually distinctive regions containing packaging text, logos, local structures, and textures, instead of uniformly averaging over low-detail product surfaces.

High-frequency discrepancy. The Fourier high-frequency metric is computed over the globally aligned whole-product region rather than individual local patches. We first take the bounding box of the aligned reference-product mask and extract the same spatial region from both the reference and generated products. Each crop is converted to luminance and multiplied by a two-dimensional Hann window before computing its Fourier spectrum. The normalized high-frequency discrepancy is defined as

$$E_{\mathrm{HF}} = \sqrt{\frac{\sum_{\mathbf{f}\in\Omega_{\mathrm{HF}}} |F_g(\mathbf{f}) - F_r(\mathbf{f})|^2}{\sum_{\mathbf{f}\in\Omega_{\mathrm{HF}}} |F_r(\mathbf{f})|^2 + \epsilon}}, \tag{31}$$

where F_r and F_g denote the Fourier spectra of the reference and generated products, respectively. We define the normalized radial high-frequency band as $\Omega_{\mathrm{HF}} = \{\mathbf{f} \mid 0.5 \leq r(\mathbf{f}) \leq 1.0\}$. A lower value therefore indicates better preservation of high-frequency product details.

Poster Text Score. Poster Text Score evaluates the accuracy of promotional copy in the generated poster. We apply PP-OCRv5 Cui et al. (2025) to the complete generated poster and match the recognized strings against the annotated promotional text using the same normalized string-similarity and one-to-one matching protocol as Product Text Score. The final score is averaged over the annotated promotional text strings.

Poster FID. We evaluate overall poster quality using Fréchet Inception Distance (FID) between the complete generated posters and their corresponding ground-truth posters. Images are resized to 299×299, and features are extracted using an ImageNet-pretrained Inception-v3 Szegedy et al. (2016) network. FID is then computed from the feature distributions over all 800 Fine-grained Fidelity samples, where a lower value indicates better agreement with the ground-truth poster distribution.

Metric scope. Product Text Score is reported on the Text-Fidelity subset, while SSIM, GMSD, Color, and HF are reported on the Visual-Fidelity subset. Poster Text Score and FID are evaluated on all 800 Fine-grained Fidelity samples and do not depend on product-detection coverage. All metric-specific preprocessing and evaluation configurations are shared across compared methods and remain fixed throughout evaluation.

E.2 Product Detection Coverage in Fine-grained Fidelity Evaluation

Table 7: Product-detection coverage for the detection-dependent subsets of the Fine-grained Fidelity evaluation. We report the detection failure rate and the number of valid samples retained for metric computation.

Method	Text-Fidelity (400)		Visual-Fidelity (529)	
	Fail. (%)	Valid	Fail. (%)	Valid
Qwen-Image-Edit	0.50	398	0.57	526
BAGEL	2.00	392	0.57	526
FLUX.2	0.25	399	0.57	526
IC-Custom	0.50	398	0.57	526
PosterOmni	0.75	397	0.38	527
OmniGen2	0.25	399	0.76	525
Step1X-Edit	1.25	395	2.65	515
FidPoster	**0.00**	**400**	**0.19**	**528**

E.3 Human Evaluation Protocol and Statistical Analysis

We conduct human evaluation on the 200 samples in the *Independent Reference* setting of FidPoster-Bench. Five annotators independently evaluate the generated results from all compared methods. For each sample, model identities are hidden and the generated posters are presented in randomized order to reduce potential evaluation bias. Each annotator assigns an integer score from 1 to 5 along four criteria: *Product Fidelity*, *Layout and Composition*, *Visual Quality*, and *Commercial Usability*. Consequently, each method receives 1,000 ratings for each criterion.

Rating criteria. Product Fidelity evaluates how faithfully the generated poster preserves distinctive characteristics of the reference product, including packaging text, logos, colors, textures, shapes, and local structures. Layout and Composition evaluates product placement, scale, spacing, visual hierarchy, and overall compositional balance. Visual Quality evaluates visual coherence, naturalness, stylistic consistency, and visible generation artifacts. Commercial Usability evaluates whether the generated poster can be practically used for commercial promotion without substantial manual correction. For all criteria, a score of 1 indicates severe deficiencies, a score of 3 indicates an acceptable result with noticeable imperfections, and a score of 5 indicates high-quality performance with no obvious correction required. Scores of 2 and 4 represent intermediate levels between these anchors.

Paired statistical analysis. The scores reported in the main paper are averages over all five annotators. For statistical analysis, we first average the five ratings for each sample–method pair, rather than treating individual annotator ratings as independent observations. This results in 200 paired sample-level scores for each criterion when comparing FidPoster with Qwen-Image-Edit, the strongest baseline in overall human evaluation.

For each criterion, let $\bar{s}_i^{\text{ours}}$ and $\bar{s}_i^{\text{base}}$ denote the mean scores of the five annotators for the i-th sample. We compute the paired difference as

$$\Delta_i = \bar{s}_i^{\text{ours}} - \bar{s}_i^{\text{base}}, \tag{32}$$

and report the mean improvement $\Delta = \frac{1}{N}\sum_{i=1}^{N} \Delta_i$, where $N = 200$.

We estimate the 95% confidence interval of Δ using 10,000 paired bootstrap resamples over the 200 test samples. At each bootstrap iteration, 200 sample pairs are drawn with replacement and their mean score difference is recomputed; the 2.5th and 97.5th percentiles of the resulting bootstrap distribution define the reported confidence interval. Statistical significance is additionally evaluated using a two-sided Wilcoxon signed-rank test on the 200 paired sample-level differences, with Holm correction across the five reported comparisons.

We also report sample-level win/tie/loss (W/T/L) rates. A sample is counted as a win if FidPoster has a higher five-annotator mean score than Qwen-Image-Edit, a tie if the two scores are equal, and a loss otherwise.

Table 8: Paired statistical analysis of human evaluation between FidPoster and Qwen-Image-Edit on the 200 Independent Reference samples. Δ denotes FidPoster $-$ Qwen-Image-Edit.

Criterion	Qwen	FidPoster	Δ (95% CI)	W/T/L (%)	p_{adj}
Product Fidelity	4.643	**4.838**	+0.195 [0.143, 0.249]	57.0 / 27.5 / 15.5	3.54×10^{-11}
Layout & Composition	4.111	**4.326**	+0.215 [0.136, 0.296]	55.5 / 16.0 / 28.5	2.58×10^{-6}
Visual Quality	4.009	**4.244**	+0.235 [0.141, 0.327]	57.5 / 14.0 / 28.5	2.58×10^{-6}
Commercial Usability	3.895	**4.235**	+0.340 [0.235, 0.446]	64.5 / 11.0 / 24.5	1.68×10^{-8}
Overall	4.165	**4.411**	+0.246 [0.171, 0.323]	**70.0 / 2.5 / 27.5**	1.28×10^{-8}

As shown in Table 8, FidPoster obtains positive mean improvements across all four evaluation criteria, and every 95% confidence interval lies entirely above zero. The largest gain is observed for Commercial Usability, with an average improvement of 0.340 and a 95% confidence interval of $[0.235, 0.446]$. For Overall score, FidPoster improves over Qwen-Image-Edit by 0.246 points and achieves a higher sample-level score on 70.0% of the test samples. These results indicate that the

human-evaluation gains are broadly distributed across test samples rather than being driven by a small number of cases.

Inter-rater agreement. We further evaluate the consistency of the five annotators using the intraclass correlation coefficient (ICC). Agreement is computed over all 1,600 sample–method pairs (200 samples $\times$ 8 methods) using a two-way random-effects, absolute-agreement formulation. We report both ICC(2,1), which measures the reliability of an individual annotator, and ICC(2,5), which measures the reliability of the mean score aggregated over all five annotators.

Table 9: Inter-rater agreement of the human evaluation. Higher ICC values indicate stronger agreement among annotators.

Criterion	**ICC(2,1)**	**ICC(2,5)**
Product Fidelity	0.624	0.892
Layout & Composition	0.627	0.894
Visual Quality	0.648	0.902
Commercial Usability	0.633	0.896
Overall	0.708	**0.924**

The aggregated five-annotator scores show consistently high agreement, with ICC(2,5) ranging from 0.892 to 0.902 across the four individual criteria and reaching 0.924 for the Overall score. This indicates that the reported human evaluation results are stable across annotators.

E.4 Inference Efficiency

Table 10: Inference efficiency at 1024×1024 resolution with 20 sampling steps, averaged over 150 images. Relative time is normalized to Qwen-Image-Edit.

Method	**Time (s/img)**$\downarrow$	**Relative Time**$\downarrow$
Qwen-Image-Edit	**80.27**	**1.00**$\times$
FidPoster	85.14	1.06$\times$