METICULOUS MAGNIFIER: EVALUATING FINE-GRAINED SUBJECT-DETAIL PRESERVATION IN SUBJECT-DRIVEN GENERATION.

Anonymous authors
Paper under double-blind review

ABSTRACT

Subject-driven generation (SDG) achieves strong high-level semantic consistency, yet preserving and evaluating localized details such as text and logos remains challenging. We study this gap by introducing DreamBench-Detail, a benchmark of detail-rich product subjects, and a controlled protocol that induces ordered levels of localized detail degradation while holding generation settings fixed. Under this protocol, embedding-based metrics show weak score separation, while holistic MLLM evaluators assign tied scores to approximately 75% of comparisons. When contextual variation is introduced, the evaluated metrics either remain weak or lose ranking accuracy. To address these limitations, we propose Meticulous Magnifier, which identifies reference-specific salient targets, localizes them in both images, and aggregates rubric-guided crop-level judgments. Across multiple distortion levels and SDG models, Meticulous Magnifier achieves 78.54% of Pairwise Accuracy with a 6.69% tie rate. It retains 75.33% accuracy under additional contextual variation, substantially outperforming the evaluated baselines.

1 INTRODUCTION

Subject-driven generation (SDG) aims to place a reference subject in a novel scene while preserving the visual attributes that make the particular instance recognizable (Ruiz et al., 2023; Gal et al., 2023; Kumari et al., 2023; Chae et al., 2023; Jun et al., 2026; Li et al., 2023a; Nam et al., 2024; Shin et al., 2025b; Tan et al., 2025). It underpins a range of practical applications, such as producing varied in-use product shots or studio-style advertisements from a single product reference. Existing evaluation protocols for general text-to-image (T2I) and image-to-image (I2I) generation largely focus on prompt adherence, perceptual quality, and broad content preservation (Hu et al., 2023; Ghosh et al., 2023; Huang et al., 2023; Zhang et al., 2023; Ku et al., 2024a;b). SDG, however, introduces an additional requirement: preserving the localized, instance-specific details that distinguish the reference subject from visually similar subjects. An output may therefore satisfy the target prompt and appear visually compelling yet remain unsuitable for applications requiring faithful reproduction when details such as text, logos, patterns, or small structural components are rendered incorrectly.

Despite substantial progress in preserving the global appearance of reference subjects, current SDG models can still distort or omit these localized attributes. Such failures are particularly consequential for detail-rich subjects, whose identity depends on the accurate reproduction of multiple fine-grained features. Recent methods have therefore introduced dedicated conditioning and refinement mechanisms to improve local subject fidelity (Jun et al., 2026; Guo & Yu, 2026; Wang et al., 2023; Shin et al., 2025a). Assessing these targeted improvements, however, requires metrics that are sensitive to localized preservation failures rather than only overall subject resemblance.

As illustrated in Fig. 1, measuring fine-grained subject preservation is challenging because reference and generated images differ both in intended ways, such as scene, viewpoint, pose, and lighting, and in unintended ways caused by failures to preserve the subject. The intended changes break the spatial correspondence assumed by full-reference metrics such as SSIM and LPIPS (Wang et al., 2004; Zhang et al., 2018), making them unsuitable for directly measuring subject consistency. To relax this requirement, early SDG works (Gal et al., 2023; Ruiz et al., 2023; Kumari et al., 2023) adopted similarities in representation spaces such as CLIP and DINO. These metrics (Radford et al., 2021;

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.

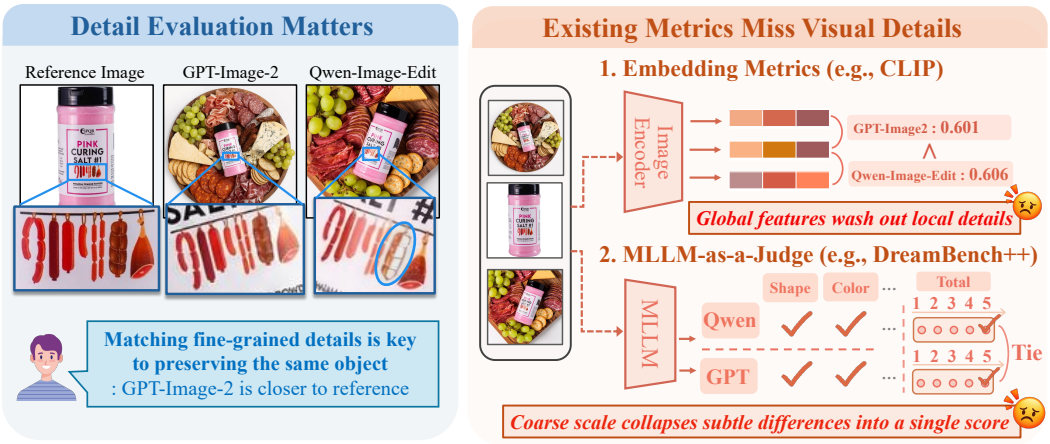

Figure 1: Given a reference and two subject-driven generated images, a human can clearly identify GPT-Image-2 as the closer match by inspecting fine-grained details (left), yet existing metrics cannot. Embedding-based metrics such as CLIP assign nearly identical scores to both outputs, and MLLM-based judges such as DreamBench++ collapse subtle differences into tied scores (right).

Oquab et al., 2023) capture broad semantic and visual correspondence, but their global representations may dilute errors confined to small, instance-specific regions. More recent evaluation methods employ MLLMs as judges (Peng et al., 2025; Ku et al., 2024a), leveraging their visual reasoning capabilities to produce assessments that align more closely with human judgment. Yet existing protocols typically produce holistic judgments over entire images, leaving unclear whether they reliably attend to and distinguish defects in fine-grained subject-details.

Studying fine-grained subject preservation requires detail-rich subjects and an evaluation setting that isolates detail differences from other visual variation. Existing SDG benchmarks primarily target overall subject consistency and offer limited coverage of localized, instance-specific attributes. We therefore construct DreamBench-Detail, using the Amazon Reviews dataset (Hou et al., 2024), comprising 439 subjects across 25 product categories and 643 reference–prompt pairs, with a controlled protocol for varying subject-detail preservation. Under this protocol, embedding-based metrics exhibit limited sensitivity to detail degradation, while holistic MLLM evaluators often fail to distinguish fine-grained differences, assigning tied scores.

To address these limitations, we propose Meticulous Magnifier, a detail-aware evaluation framework that replaces a single holistic judgment with reference-specific local assessments. By identifying salient visual targets, grounding each comparison in the corresponding image regions, and aggregating rubric-guided target-level judgments, Meticulous Magnifier provides a finer-grained measure of subject detail preservation. It achieves 78.54% pairwise accuracy with a 6.69% tie rate and retains its advantage when additional contextual variation is introduced.

We summarize our main contributions as follows:

- We introduce DreamBench-Detail, a human-verified benchmark comprising 439 detail-rich product subjects across 25 categories and 643 reference–prompt pairs, together with a controlled protocol for evaluating metric sensitivity to localized subject-detail degradation.
- Using this protocol, we show that commonly used embedding-based metrics exhibit limited sensitivity to fine-grained detail changes, while holistic MLLM evaluators fail to distinguish many such differences, assigning tied scores to about 75% of pairwise comparisons.
- To address these limitations, we propose Meticulous Magnifier, which replaces holistic comparison with localized, rubric-guided judgments of reference-specific details. It achieves 78.54% intervention-consistent pairwise accuracy with a 6.69% tie rate and remains effective under additional contextual variation.

2 Related Work

2.1 Subject-Driven Generation and Detail Preservation

Early SDG approaches constructed subject-specific representations by fine-tuning on a small number of reference images. For example, Textual Inversion (Gal et al., 2023) learns a token embedding for a target subject, allowing the subject to be invoked through a unique token in text prompts. DreamBooth (Ruiz et al., 2023) instead fine-tunes the generative model to associate a unique identifier with the subject's appearance. Although effective, these approaches require per-subject adaptation.

Subsequent methods avoid per-subject optimization by directly conditioning generation on reference images. BLIP-Diffusion (Li et al., 2023a), InstantBooth (Shi et al., 2024), and Subject-Diffusion (Ma et al., 2024) learn generalizable subject representations from reference images. IP-Adapter (Ye et al., 2023) introduces a dedicated image-conditioning pathway alongside text conditioning, whereas Kosmos-G (Pan et al., 2024) uses an MLLM to encode interleaved image–text inputs for image generation. Other approaches introduce more explicit mechanisms for preserving subject appearance. AnyDoor (Chen et al., 2024b) combines identity and detail features for zero-shot object-level customization, DreamMatcher (Nam et al., 2024) matches semantically corresponding features between the reference and generated images, and OminiControl (Tan et al., 2025) provides parameter-efficient image conditioning for diffusion transformers. More recently, foundation models such as FLUX.1-Kontext (Labs et al., 2025) and Qwen-Image-Edit (Wu et al., 2025) have enabled in-context generation and editing directly from reference images without per-subject optimization.

Despite these advances, preserving the global appearance of a subject does not guarantee faithful reproduction of localized attributes such as text, logos, patterns, textures, and small structural components. FlowFixer (Jun et al., 2026), for example, introduces a post-generation refinement model for restoring subject details. Whereas these methods seek to improve generation or refinement, our work focuses on evaluating whether localized, instance-specific details have been preserved.

2.2 Image Similarity and Subject-Consistency Evaluation

Comparing images as humans perceive them has been a longstanding goal in computer vision. Classical metrics such as L2 distance and PSNR quantify direct pixel-level errors, while SSIM (Wang et al., 2004) emphasizes local structural similarity. LPIPS (Zhang et al., 2018) instead compares deep feature activations and better reflects human perceptual judgments than pixel-level distances. Other approaches measure similarity using representations from large vision models, including CLIP (Radford et al., 2021) and DINO (Oquab et al., 2023), or learn a perceptual similarity space directly from human judgments, as in DreamSim (Fu et al., 2023).

Evaluating subject consistency in SDG differs from conventional full-reference comparison because the subject must be assessed independently of intentional changes such as context, pose, viewpoint, and lighting. Prior SDG works therefore commonly use CLIP and DINO similarities to measure semantic and visual correspondence (Ruiz et al., 2023; Gal et al., 2023; Shin et al., 2025b). RefVNLI (Slobodkin et al., 2025) instead learns a scalable vision–language inference metric that jointly evaluates textual alignment and subject preservation. Detail-oriented methods have additionally considered localized correspondence, such as keypoint matching (Jun et al., 2026). While these approaches capture complementary aspects of subject consistency, they do not explicitly assess a reference-specific set of localized attributes. A metric may therefore recognize the same overall subject while remaining insensitive to errors in the details that distinguish the particular instance.

2.3 MLLM-as-a-Judge for Image Generation

The visual reasoning and instruction-following capabilities of MLLMs have motivated their use as evaluators of generated images. VIEScore (Ku et al., 2024a) uses an MLLM to assess conditional image synthesis and explain its judgment, while DreamBench++ (Peng et al., 2025) evaluates subject consistency using broad criteria such as shape, color, and texture. Related frameworks use visual question answering or compositional tests to determine whether generated images satisfy specified conditions (Hu et al., 2023; Huang et al., 2023).

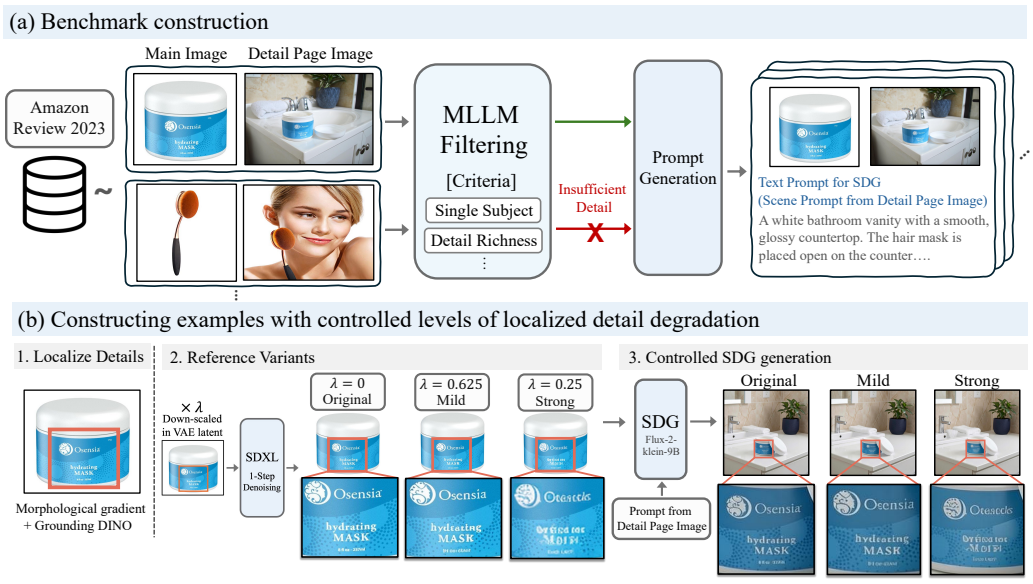

Figure 2: Overview of (a) the DreamBench-Detail construction pipeline and (b) the controlled detail degradation protocol while remaining composition largely fixed. Best viewed in zoomed.

MLLM judgments nevertheless remain sensitive to high-score bias, inconsistency, positional bias, and evaluator choice (Chen et al., 2024a; Hwang et al., 2025; Pu et al., 2025). Their visual grounding can also be unreliable, as object-hallucination studies show that MLLMs may describe content absent from an image (Li et al., 2023b; Park & Li, 2025). These limitations are particularly relevant to subject-detail evaluation. By summarizing an entire reference–generation pair with a single holistic score, existing SDG evaluators leave it unclear whether small, reference-specific regions were examined at all, and lack the resolution to capture subtle differences.

Our method instead decomposes subject identity into reference-specific visual targets, grounds each assessment in localized image evidence, and aggregates rubric-guided target-level judgments. This formulation differs from existing holistic MLLM evaluators in both the unit of evaluation and the construction of the final score. We evaluate the effects of target decomposition, localization, and score aggregation through controlled experiments and ablations.

3 Can Existing Metrics Detect Subject-Detail Differences?

We investigate whether existing SDG subject-consistency metrics can detect differences in fine-grained details. We first construct a benchmark of detail-rich subjects (Sec. 3.1) and introduce a controlled protocol that varies localized subject detail while holding other generation factors fixed (Sec. 3.2). We define the evaluation criteria (Sec. 3.3) and evaluate representative metrics (Sec. 3.4).

3.1 Benchmark Dataset

Existing SDG benchmarks range from the relatively small subject collections introduced by early personalization methods (Ruiz et al., 2023; Kumari et al., 2023) to larger benchmarks such as DreamBench++ (Peng et al., 2025). However, they primarily target overall subject consistency and provide limited coverage of detail-rich products whose identity depends on localized attributes.

To construct a benchmark of diverse subjects with salient, instance-specific details, we build DreamBench-Detail from the Amazon Reviews 2023 dataset (Hou et al., 2024). For each product, we select a representative image as the SDG reference and derive target-scene prompts from the corresponding product-detail images. Candidate references and target images are filtered by an MLLM and subsequently verified by human reviewers as shown in Fig. 2. The resulting benchmark contains 439 subjects across 25 product categories and 643 reference-prompt pairs, with one to three prompts per subject. Additional construction details are provided in Appendix A.

Table 1: Pairwise comparisons among three distortion levels. Acc. denotes Pairwise Accuracy. Emb., Corr., and MLLM denote embedding, correspondence, and MLLM based metric categories, respectively.

Category	Metric	Acc. %	Tie %	AUC_z
Emb.	CLIP	51.58	0.16	10.47
	DINO	60.24	0.05	8.34
Corr.	Keypoints	70.76	1.60	28.25
MLLM	DB++	12.34	78.80	12.34
	VIEScore	16.23	72.89	10.11
	MetMag	78.54	6.69	53.23

Table 2: Results of context variation stress test. Abbreviations follow Table 1.

Category	Metric	Acc. (%)	ΔAcc	AUC_z
Emb.	CLIP	51.87	+0.29	29.81
	DINO	51.93	−8.31	30.37
Corr.	Keypoints	61.47	−9.29	36.45
MLLM	DB++	21.80	+9.46	21.80
	VIEScore	26.33	+10.10	18.47
	MetMag	75.33	−3.21	55.83

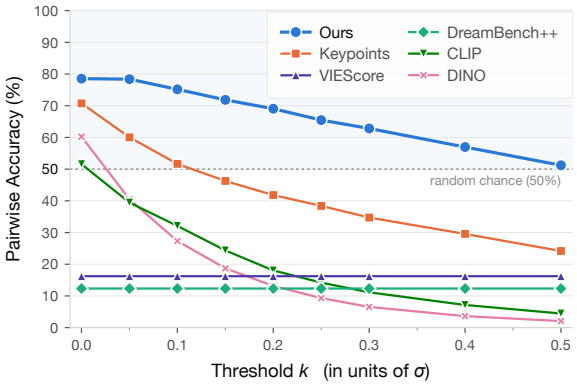

Figure 3: Pairwise Accuracy as a function of the normalized threshold k. Ours sustains substantially higher accuracy than all baselines as k increases, indicating stronger score separation. The dotted line marks random chance (50%). For visibility, the graph is trimmed at $k = 0.5$; see Appendix C for the full version.

3.2 Controlled SDG Evaluation Protocol

In standard SDG, generated outputs differ simultaneously in subject detail, composition, pose, lighting, and other visual properties, making it difficult to determine which differences drive a score. We therefore construct outputs with controlled levels of localized detail degradation.

For each reference, we apply one-step noise-denoising and vary its VAE input resolution among $1\times$, $0.625\times$, and $0.25\times$, corresponding to approximately 100%, 40%, and 6% of the original pixel area. We use a fixed timestep of 10 and restrict the intervention to localized detail regions. Full implementation details and validation of the degradation procedure are provided in Appendix B.

We generate an SDG output from each reference variant while fixing the model, prompt, random seed, and all other generation settings. This produces three conditions per prompt with increasing source-detail degradation and largely consistent composition (Fig. 2). We denote these None, Mild, and Strong according to the distortion added to their source reference. Using FLUX.2-klein-9B (Labs, 2026), the protocol yields 1,929 outputs and 1,929 pairwise comparisons across 643 prompts. We observe consistent trends with FLUX.1-Kontext (Labs et al., 2025) and Qwen-Image-Edit-2511 (Wu et al., 2025), as reported in Appendix D.

3.3 Evaluation Criteria

Under our controlled protocol, a detail-sensitive metric should assign a higher score to an output conditioned on a less-degraded reference. We measure this property using **Pairwise Accuracy**, defined as the proportion of pairs ranked in the expected order. Scores that are equal after rounding to four decimal places are treated as ties and counted as incorrect.

Since Pairwise Accuracy captures ranking correctness but not the magnitude of score separation, we also report $\mathbf{AUC}_z$. Let σ denote the standard deviation of a metric's scores over the evaluation set. At normalized margin k, a pair is treated as tie if its absolute score difference is smaller than $k\sigma$. AUC_z is the area under the resulting Pairwise Accuracy curve as k varies from 0 to 1. Normalization by σ enables comparison across metrics with different numerical scales.

3.4 Results on Existing SDG Metrics

We evaluate five representative baselines: CLIP (Radford et al., 2021), DINO (Oquab et al., 2023), VIEScore (Ku et al., 2024a), DreamBench++ (Peng et al., 2025), and matched-keypoint counts as

an absolute score, adapted from FlowFixer (Jun et al., 2026). Tab. 1 summarizes the results, with breakdowns by degradation severity provided in Appendix C.

Controlled comparison. CLIP performs near chance at 51.58% Pairwise Accuracy, while DINO reaches 60.24%. Despite DINO's higher ranking accuracy, both yield low AUC_z scores of 10.47 and 8.34, respectively, indicating weak score separation under localized detail changes. DreamBench++ and VIEScore assign tied scores to 78.80% and 72.89% of comparisons, respectively. This indicates that their holistic, discrete scoring protocols rarely separate outputs produced at different degradation levels. While keypoint matching is the most sensitive metric (70.76% Pairwise Accuracy, AUC_z of 28.25), its AUC curve drops sharply at small margins as shown in Fig. 3, showing that detail distortion induces only small score differences relative to its overall variation.

Context-variation stress test. We further compare outputs generated using different prompts and reference degradation levels, introducing contextual variation while retaining the known ordering of the detail intervention. As shown in Tab. 2, Pairwise Accuracy of DINO and keypoint matching decrease by 8.31%p and 9.29%p relative to the controlled comparison. In contrast, those of DreamBench++ and VIEScore increase to 21.80% and 26.33%, respectively, indicating that their judgments are heavily influenced by visual context rather than by subject details. Overall, the evaluated metrics provide weak score separation in the controlled setting or lose ranking accuracy when detail differences coexist with contextual variation.

4 Meticulous Magnifier: Structured Local Detail Evaluation

4.1 Motivation

The poor performance of holistic MLLM evaluators in Sec. 3 does not necessarily imply that MLLMs cannot recognize subject details. Existing protocols present the full reference-generation pair with abstract criteria and ask the evaluator to summarize their similarity using a single score. We therefore investigate whether restructuring the judging protocol around explicitly identified and localized details can improve sensitivity to subject-detail differences.

Holistic evaluation leaves three decisions coupled within a single judgment: what reference-specific details should be evaluated, where the corresponding evidence should be examined, and how multiple local judgments should be combined into a final score. Broad criteria such as shape, color, and texture do not specify which attributes distinguish the particular reference subject and where those attributes should be examined. Moreover, the high tie rates in Tab. 1 indicate that a single coarse score provides limited resolution for distinguishing subtle differences.

As shown in Fig. 4, Meticulous Magnifier explicitly decomposes those three stages: reference-specific target generation determines what to evaluate, open-vocabulary grounding and cropping determine where to examine each target, and aggregation of rubric-guided local judgments determine how to construct the final score with enough resolution.

4.2 Method

Multi-Scale Salient Target Generation. The first stage determines what to evaluate by constructing a reference-specific set of visual targets. We define a salient visual target as a visually identifiable subject component or attribute whose preservation is important for characterizing the particular reference subject. We refer to the resulting collection as the Salient Target List.

Given a reference image, an MLLM identifies targets at multiple levels of granularity, ranging from overall subject structure and major components to localized details such as text, logos, patterns, and textures. The Salient Target List is generated once per reference image and reused across all outputs evaluated against that reference. This ensures that the same reference-specific criteria are applied when comparing different outputs from the same reference.

Open-Vocabulary Grounding and Crop Extraction. The second stage determines where to examine each target. For every target in the Salient Target List, an open-vocabulary grounding model

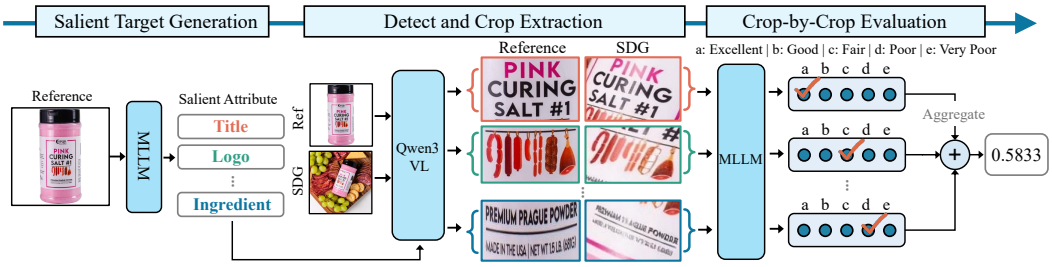

Figure 4: The pipeline of Meticulous Magnifier. An MLLM first generates a reference specific Salient Target List, each target is then localized and cropped in both the reference and generated images using open vocabulary grounding, and finally an MLLM judges each cropped pair using a five level rubric before the local scores are aggregated into a final score.

predicts a bounding box in both the reference and generated images. We crop the corresponding regions to obtain a paired local observation for that target. The reference bounding boxes are computed once and reused across evaluations, whereas the generated-image bounding boxes are predicted independently for each output. By presenting the evaluator with only these paired crops, we remove irrelevant visual context and concentrate its judgment on the relevant subject detail.

Crop-Level MLLM Judging and Score Aggregation. In the final stage, the MLLM independently evaluates each paired set of reference and generated crops. The evaluator receives only the two corresponding crops, without the target title or other target-specific text, and judges how faithfully the generated crop preserves the visual content of the reference crop.

We use a predefined five-level ordinal rubric: Very Poor, Poor, Fair, Good, and Excellent, mapped to scores of 0, 1, 2, 3, and 4, respectively. Each level is accompanied by an explicit definition, and the evaluator is instructed to consider properties such as geometry, structural consistency, and fine appearance. The MLLM also produces a concise description of the observed correspondence or discrepancy for qualitative inspection, while the categorical judgment alone serves as the basis for quantitative evaluation. The final subject-detail preservation score is computed by averaging the local scores and normalizing the result to the range [0,1]. We use Qwen3-VL-Instruct-8B (Bai et al., 2025) as the evaluator.

4.3 Comparison with Existing Metrics

Controlled comparison. As shown in Tab. 1, Meticulous Magnifier achieves 78.54% Pairwise Accuracy, outperforming the strongest existing baseline, keypoint matching, by 7.78 percentage points. It also achieves the highest AUC_z of 53.23, compared with 28.25 for keypoint matching, indicating greater score separation as larger normalized margins are required. Although each local judgment is itself a five-level ordinal decision, aggregating multiple reference-specific judgments yields a tie rate of only 6.69%, compared with 78.80% for DreamBench++ and 72.89% for VIEScore.

Context-variation stress test. When outputs differ in both prompt context and reference-degradation level, Meticulous Magnifier achieves 75.33% Pairwise Accuracy, a decrease of 3.21 percentage points from the controlled comparison, as shown in Tab. 2. This remains the highest accuracy among the evaluated metrics, with a smaller decrease than those observed for DINO and keypoint matching. Thus, Meticulous Magnifier retains most of its controlled ranking performance when localized detail differences coexist with contextual variation.

5 Analysis and Application

5.1 Design Choices for MLLM Usage

We ablate our three-stage design (Sec. 4.2) by progressively removing its components, moving from fully holistic evaluation to our complete localized pipeline in Tab. 3. **Holistic (w/o criteria)** and **Holistic (w/ criteria)** correspond to VIEScore and DreamBench++, respectively, which prompt the

Table 3: Ablation on MLLM judging design, comparing holistic evaluation against progressively added components of our per target rubric pipeline.

Design	Pairwise Accuracy (%)	Tie Rate (%)	AUC_z
Holistic evaluation			
w/o criteria (VIEScore)	16.23	72.89	10.11
w/ criteria (DreamBench++)	12.34	78.80	12.34
Per-target rubric evaluation			
w/o localization (Ours-aggregate)	74.70	12.60	46.47
+ bbox, w/o crop	72.63	10.26	43.57
+ bbox + crop (Ours)	**78.54**	**6.69**	**53.23**

MLLM to compare the images holistically and produce a single similarity score, without or with broad evaluation criteria (e.g., shape, color, and texture). **Explicit rubric (w/o localization)**, denoted Ours-aggregate, replaces these broad criteria with our per-target Salient Target List and aggregates judgments on each target. **Rubric + bbox (w/o crop)** additionally localizes each target with a bounding box, but overlays it on the full image rather than cropping, so the MLLM still sees the entire image alongside the highlighted region. **Rubric + bbox + crop** is our full method, which crops each localized region before evaluation.

The largest single gain comes from replacing holistic evaluation with an explicit, reference-specific rubric and score aggregation. Pairwise Accuracy rises from the 12–16% of holistic methods to 74.7%, and the tie rate drops from over 70% to 12.6%. This shows that the primary bottleneck is the abstract criteria and coarse scoring of holistic MLLM-as-a-judge metrics, not the MLLM's visual reasoning ability. Moreover, adding a bounding box without cropping (Rubric + bbox, w/o crop) does not help and slightly hurts performance (72.63% vs. 74.70% for Ours-aggregate), likely because overlaying a box on the full image distracts the model rather than focusing it on the relevant region. Only cropping and isolating the target region yields a clear benefit, confirming that cropping, not localization information alone, drives the gain from our target-grounding stage.

Table 4: Ablation on the granularity of the Salient Target List.

Granularity	Acc. (%)	Tie (%)	AUC_z
Coarse-only	71.85	16.59	51.56
Fine-only	77.86	8.35	**54.24**
Mixed (Ours)	**78.54**	**6.69**	53.23

Table 5: Ablation on underlying MLLM used for crop level judging, varying model size and model family.

Category	Model	Acc. (%)	AUC_z
Baseline	Qwen3-VL-8B	78.54	53.23
Model Size	Qwen3-VL-32B	79.57	54.42
Model Family (8B scale)	GLM-4.6V-Flash	73.72	51.18
	InternVL3-8B	74.55	42.91

5.2 Salient Target List Granularity

Tab. 4 presents an ablation of the multi-scale target representation, in which the Salient Target List is restricted to a single granularity level. Restricting the targets to coarse functional regions (Coarse-only) more than doubles the tie rate (16.59% vs. 6.69% for Mixed), because a coarse region spans a broad area over which subtle detail distortions can be averaged out or judged unchanged. Restricting the targets to fine-grained components (Fine-only) recovers most of this gap, reaching 77.86% Pairwise Accuracy and a marginally higher AUC_z than Mixed (54.24 vs. 53.23), which confirms that fine-grained targets carry most of the detail-sensitivity signal. Fine-only nonetheless yields a higher tie rate (8.35% vs. 6.69%), indicating that coarse targets, such as the relationships among multiple attributes, also contribute to the sensitive evaluation of detail preservation. Combining the two (Mixed, Ours) achieves the highest Pairwise Accuracy and the lowest tie rate while retaining score separation comparable to Fine-only, confirming that the two target types are complementary.

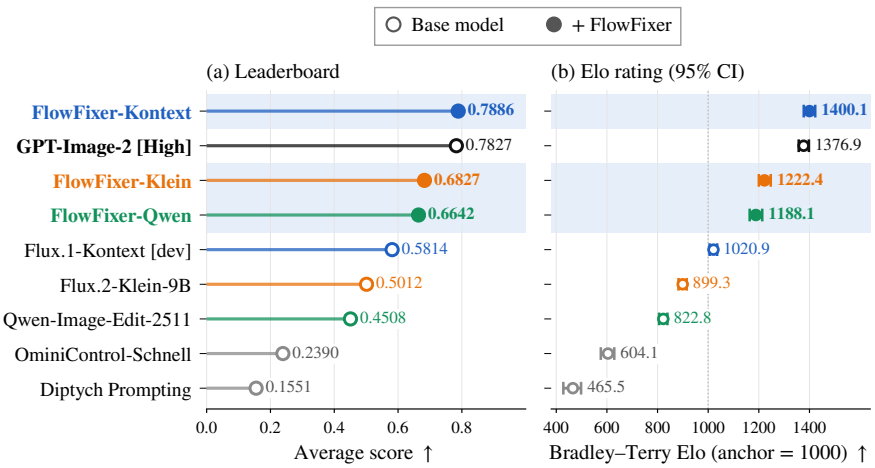

Figure 5: (a) Leaderboard and (b) Elo ratings of representative SDG models on our benchmark.

5.3 MLLM Dependency

In Tab. 5, we examine whether our framework's effectiveness depends on a specific underlying MLLM, varying both model scale and model family. Scaling the underlying MLLM from 8B to 32B parameters yields only a modest improvement (78.54% $\rightarrow$ 79.57% pairwise accuracy, 53.23 $\rightarrow$ 54.42 AUC_z), indicating that our framework does not rely on a particularly large model to be effective. Across model families, all three 8B-scale MLLMs achieve strict Pairwise Accuracy in the 73-79% range with tie rates below 8%, substantially outperforming every existing metric evaluated in Sec. 3.4, though some variation in AUC_z remains. Together, these results suggest that the gains observed in this work stem primarily from the proposed decoupled evaluation architecture rather than from the raw capability of any single MLLM backbone.

5.4 Application: SDG Model Comparison

Using Meticulous Magnifier scores on DreamBench-Detail, we present a leaderboard and Elo ratings for representative SDG models, including our detail refinement method FlowFixer, in Fig. 5. The results yield a few notable findings. First, without refinement, the performance gap between the proprietary model (GPT-Image-2[High]) and open-source baseline models is substantial. Second, more recent models do not always outperform earlier ones in subject-detail preservation; for example, Flux.1-Kontext[dev] ranks higher than Flux.2-Klein-9B. Finally, refining details of generated images brings a substantial gain, where FlowFixer applied on images generated with Flux.1-Kontext[dev] narrowly surpasses GPT-Image-2[High] in subject-detail preservation capability, consistently across both the leaderboard and Elo ratings.

6 Conclusion

In this work, we ask whether existing subject-consistency metrics can reliably measure fine-grained subject-detail preservation in subject-driven generation (SDG). Using DreamBench-Detail and a controlled protocol that varies only subject detail, we evaluated metrics along two axes, discernment (Pairwise Accuracy) and sensitivity (AUC_z), and found that both embedding-similarity and MLLM-as-a-judge metrics are insufficiently sensitive to fine details of the subject. Motivated by this analysis, we proposed Meticulous Magnifier that identifies a multi-scale set of salient targets, localizes and crops each one, and aggregates independent crop-level judgments. This metric attains substantially higher discernment and sensitivity than existing metrics and remains robust even when visual context varies, and our ablations show that all components of the pipeline are essential to this improvement. However, Meticulous Magnifier still relies on coarse per-crop scoring, and these scores may become inflated as the overall performance of SDG models improves substantially. Moreover, because Meticulous Magnifier measures only subject-detail preservation, a complete assessment of SDG performance should jointly consider complementary criteria such as prompt fidelity and aesthetic quality. We leave finer score resolution and a unified metric to future work.

7 AI USE STATEMENT

In this work, we used generative AI tools only for language editing, including grammar correction, rephrasing, and improving the clarity of the manuscript. Additionally, we used generative AI tools to assist with code implementation and finding related works. All AI-generated code was reviewed and tested for correctness by the authors. We have not used generative AI tools for research ideation, method or experiment design, data analysis, or the generation of scientific claims, and other required-disclosure tasks are not applicable to this work. We have reviewed all AI-assisted edits: every suggestion was manually checked by the authors to ensure it preserved the intended technical meaning, and no text was included without author verification. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

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

Daewon Chae, Nokyung Park, Jinkyu Kim, and Kimin Lee. InstructBooth: Instruction-following personalized text-to-image generation. *arXiv preprint arXiv:2312.03011*, 2023.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. MLLM-as-a-judge: Assessing multimodal LLM-as-a-judge with vision-language benchmark. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 6562–6595, 2024a.

Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. AnyDoor: Zero-shot object-level image customization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6593–6602, 2024b.

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. DreamSim: Learning new dimensions of human visual similarity using synthetic data. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using Textual Inversion. In *The Eleventh International Conference on Learning Representations*, 2023.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. GenEval: An object-focused framework for evaluating text-to-image alignment. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Hanzhong Guo and Yizhou Yu. Decomposing subject-driven image generation via intermediate structural prediction. *arXiv preprint arXiv:2605.20807*, 2026.

Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. Bridging language and items for retrieval and recommendation. *arXiv preprint arXiv:2403.03952*, 2024.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A. Smith. TIFA: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20406–20417, 2023.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2I-CompBench: A comprehensive benchmark for open-world compositional text-to-image generation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 78723–78747, 2023.

Yerin Hwang, Dongryeol Lee, Kyungmin Min, Taegwan Kang, Yongil Kim, and Kyomin Jung. Fooling the LVLM judges: Visual biases in LVLM-based evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 23197–23216, 2025.

Jinyoung Jun, Won-Dong Jang, Wenbin Ouyang, Raghudeep Gadde, and Jungbeom Lee. FlowFixer: Towards detail-preserving subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

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

Max Ku, Tianle Li, Kai Zhang, Yujie Lu, Xingyu Fu, Wenwen Zhuang, and Wenhu Chen. ImagenHub: Standardizing the evaluation of conditional image generation models. In *The Twelfth International Conference on Learning Representations*, 2024b.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1931–1941, 2023.

Black Forest Labs. FLUX.2 [Klein]: Towards interactive visual intelligence, 2026. URL `https://bfl.ai/blog/flux2-klein-towards-interactive-visual-intelligence`.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. FLUX.1 Kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Dongxu Li, Junnan Li, and Steven C. H. Hoi. BLIP-Diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. In *Advances in Neural Information Processing Systems*, volume 36, pp. 30146–30166, 2023a.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305, 2023b.

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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 38–55, 2024.

Jian Ma, Junhao Liang, Chen Chen, and Haonan Lu. Subject-Diffusion: Open domain personalized text-to-image generation without test-time fine-tuning. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–12, 2024.

Jisu Nam, Heesu Kim, DongJae Lee, Siyoon Jin, Seungryong Kim, and Seunggyu Chang. DreamMatcher: Appearance matching self-attention for semantically-consistent text-to-image personalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8100–8110, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Xichen Pan, Li Dong, Shaohan Huang, Zhiliang Peng, Wenhu Chen, and Furu Wei. Kosmos-G: Generating images in context with multimodal large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

Seongheon Park and Sharon Li. GLSim: Detecting object hallucinations in LVLMs via global-local similarity. In *Advances in Neural Information Processing Systems*, volume 38, pp. 37747–37776, 2025.

Yuang Peng, Yuxin Cui, Haomiao Tang, Zekun Qi, Runpei Dong, Jing Bai, Zheng Ge, Xiangyu Zhang, Shu-Tao Xia, et al. DreamBench++: A human-aligned benchmark for personalized image generation. In *The Thirteenth International Conference on Learning Representations*, 2025.

Shu Pu, Yaochen Wang, Dongping Chen, Yuhang Chen, Guohao Wang, Qi Qin, Zhongyi Zhang, Zhiyuan Zhang, Zetong Zhou, Shuang Gong, et al. Judge anything: MLLM as a judge across any modality. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 5742–5753, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763, 2021.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. DreamBooth: Fine-tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22500–22510, 2023.

Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. InstantBooth: Personalized text-to-image generation without test-time finetuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8543–8552, 2024.

Chaehun Shin, Jooyoung Choi, Johan Barthelemy, Jungbeom Lee, and Sungroh Yoon. Negative-guided subject fidelity optimization for zero-shot subject-driven generation. *arXiv preprint arXiv:2506.03621*, 2025a.

Chaehun Shin, Jooyoung Choi, Heeseung Kim, and Sungroh Yoon. Large-scale text-to-image model with inpainting is a zero-shot subject-driven image generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7986–7996, 2025b.

Aviv Slobodkin, Hagai Taitelbaum, Yonatan Bitton, Brian Gordon, Michal Sokolik, Nitzan Bitton Guetta, Almog Gueta, Royi Rassin, Dani Lischinski, and Idan Szpektor. RefVNLI: Towards scalable evaluation of subject-driven text-to-image generation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 8420–8438, 2025.

Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. OminiControl: Minimal and universal control for diffusion transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14940–14950, 2025.

Zhonghao Wang, Wei Wei, Yang Zhao, Zhisheng Xiao, Mark Hasegawa-Johnson, Humphrey Shi, and Tingbo Hou. HiFi Tuner: High-fidelity subject-driven fine-tuning for diffusion models. *arXiv preprint arXiv:2312.00079*, 2023.

Zhou Wang, Alan Bovik, H.R. Sheikh, and Eero Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

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.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. IP-Adapter: Text-compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

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

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, 2018.

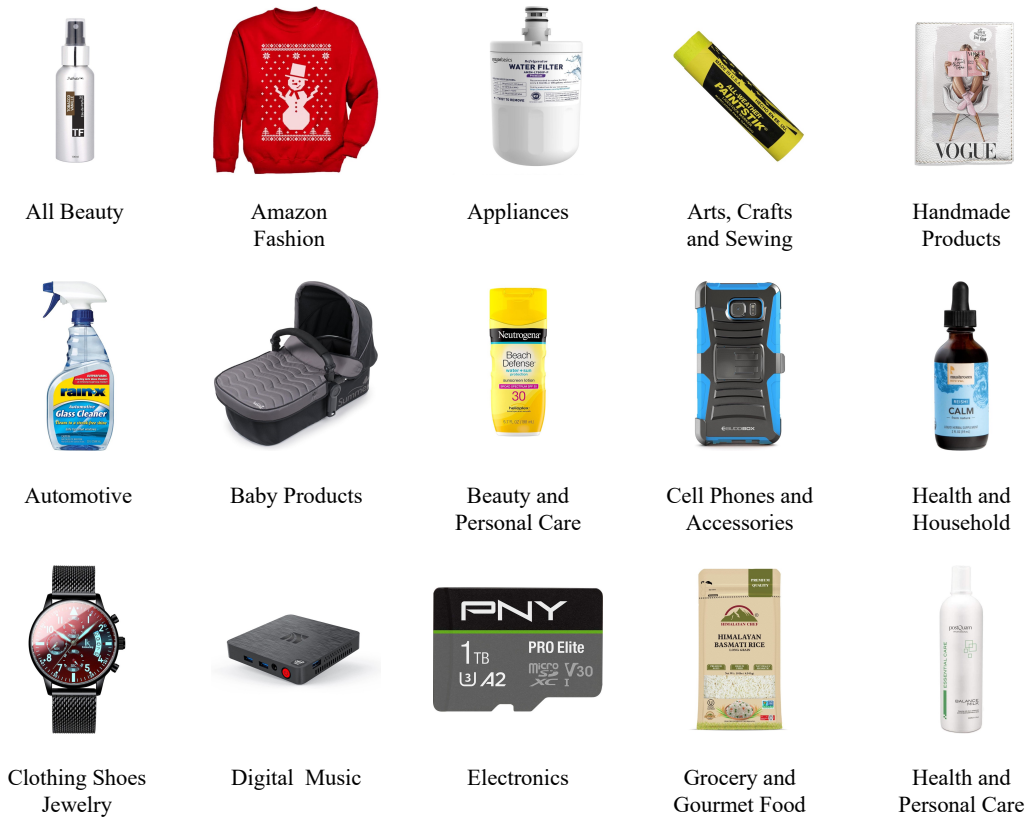

Figure 6: Visualizations of DreamBench-Detail per categories.

A DreamBench-Detail construction pipeline

To construct DreamBench-Detail, we collect product images from the Amazon Reviews 2023 dataset (Hou et al., 2024) and remove products composed of multiple components, together with product images that contain extraneous objects or distracting background elements. To extract SDG text prompts from product detail-page images, we present both the reference image and the product detail-page image to an MLLM and instruct it to describe the context of the target image in detail. To prevent the prompt from revealing which details should be preserved from the reference image, we require the model to refer to the product only by a generic category description inferred from its Amazon product title (for example, "hair mask" rather than the full product name), and to omit any subject-specific visual details from the prompt. Fig. 7 and 8 present the text prompts used for filtering.

We employ Qwen3-VL-32B-Instruct to automate this filtering process. For quality control, we subsequently inspect the collected data manually to verify that these instructions were followed. Examples from various categories are shown in Fig. 6.

You are a data curator for a reference-based commercial image generation dataset.
Your task is to determine whether a product image is suitable as a **REFERENCE IMAGE** for subject-driven image generation.
You will receive:

- Product Name
- Candidate Reference Image

Critical Rule. All decisions must be based strictly on direct visual evidence present in the Candidate Reference Image. Do not make assumptions or decisions based on common

knowledge about the product, nor be influenced by the Product Name (title). The Product Name is provided solely to help identify the category and location of the subject in the image, but the final evaluation of `ISOLATION` and `DETAIL_RICHNESS` must rely entirely on what is visually visible.
A valid reference image must satisfy both criteria.

1. `ISOLATION`
2. `DETAIL_RICHNESS`

Isolation Determine whether the image provides a clear and unambiguous visual reference of a single, coherent product subject suitable for subject-driven image generation. Use the Product Name to identify the intended product category and subject.
PASS if and only if ALL of the following conditions are simultaneously met.

- The image contains exactly ONE, SINGLE item. No multiple items even though it is a set of same items.
- No other items are in the image except the subject item.
- The subject is composed of a single object.
- The subject is clearly identifiable in the image.
- The subject is not severely cropped or occluded in the image.
- The subject is not a multi-part kit, bundle, or assembly of loose, heterogeneous items.
- If the subject is a cloth or accessory, it must be laid flat on a plain background.

Detail Richness To make exact subject preservation meaningful and non-trivial, determine whether the visible subject contains complex, fine-grained visual details (beyond simple, low-complexity features or generic category-level characteristics). Judge the subject itself in the image, not the complexity or quality of the image.
Identity-bearing details are complex, fine-grained visual characteristics that distinguish this specific product from other products in the same category. Simple, low-complexity features (even if unique) or generic category-level features do not qualify.
Check for the following.

- Complex shape or intricate local geometry (not simple shapes like a basic cylinder, cube, or slight curve).
- Detailed structural parts or components (e.g., complex stitching, buckles, or intricate assembly parts).
- Fine-grained materials, textures, or surface finishes (e.g., detailed wood grain, knit patterns, embossed textures, or brush strokes).
- Complex patterns, prints, or decorative elements (not simple solid color blocks or basic geometric patterns).
- Highly specific, detailed logos, labels, typography, or graphics.
- Intricate spatial arrangement of these elements that makes exact subject preservation non-trivial.

PASS if the subject has multiple fine-grained, complex details that establish its exact identity beyond its category and make subject preservation non-trivial.
FAIL if the subject is visually plain, generic, or lacks distinguishing details (e.g., a solid-colored t-shirt or basic plain bottle).

Final Decision **ACCEPT** if and only if

- `ISOLATION` = PASS
- `DETAIL_RICHNESS` = PASS

Otherwise, **REJECT**.

Output Format Output exactly one valid JSON object.

```
{
  "isolation": "PASS" | "FAIL",
  "detail_richness": "PASS" | "FAIL",
  "decision": "ACCEPT" | "REJECT"
}
```

Rules:

- The final decision must follow the stated rule exactly.
- Output ONLY the JSON object. Do not output markdown, reasoning, or additional text.

Figure 7: Subject-Driven Generation Reference Validity Prompt.

You are a data curator evaluating if a Candidate Target Image is a valid generation target for a given Reference Image in a subject-driven image generation dataset.
Both criteria below must PASS for a final decision of ACCEPT. Otherwise, REJECT.
You will receive:

- Reference Image (Image 1)
- Candidate Variation Image (Image 2)

Subject Validity Determine if the target subject matches the exact visual identity of the reference subject. Rely only on visual cues, not listing metadata or external knowledge.
PASS if:

- The target shows the same visual subject (allowing natural changes in viewpoint, scale, pose, lighting, or partial occlusion).

FAIL if:

- The reference subject is absent or target shows a different subject.
- The target isolates a component of the reference assembly as the main subject.
- The target requires product details not visible from the reference (i.e. generating the target requires inventing unseen details).

Test. "Could the target's identity-bearing details be generated solely from the visual information in the reference?" (PASS only if yes)

Failure Reasons (select exactly one if FAIL, else null):

- `SUBJECT_ABSENT`: Reference subject is not in target.
- `DIFFERENT_SUBJECT`: Target shows a different product.
- `CATEGORY_MATCH_ONLY`: Only a similar category-level product is shown.
- `COMPONENT_PROMOTED_TO_SUBJECT`: A component of the reference assembly is framed as the main subject.
- `UNSEEN_IDENTITY_INFORMATION`: Target relies on unseen identity details.
- `SUBJECT_NOT_IDENTIFIABLE`: Cannot identify the subject in the target.

SDG Appropriateness Determine if the target represents a meaningful subject-driven generation task (e.g., placing the subject in a new scene/context), not a simple view change or graphic.
PASS if:

- The target introduces a meaningful change (e.g., new environment, context, real-world interaction, or physical configuration).

FAIL if the main difference is only:

- Viewpoint/framing changes (viewpoint, crop, zoom, close-up, or minor lighting/color changes).
- Component focus (magnified detail or focus on one component).
- Graphic presentations (infographics, diagrams, text/graphics overlays, collages, or split-screen layouts).

Test. "Would a natural generation prompt request a new scene/interaction of the subject (PASS), or merely another view, closer inspection, or graphic layout (FAIL)?"

Failure Reasons (select exactly one if FAIL, else null):

- `VIEWPOINT_ONLY`: Only viewpoint/rotation is changed.
- `CROP_OR_ZOOM_ONLY`: Only framing/crop/zoom is changed.
- `DETAIL_CLOSEUP_ONLY`: Only a close-up/magnified detail is shown.
- `COMPONENT_FOCUS_ONLY`: Focus shifts to a single component.
- `LIGHTING_OR_COLOR_ONLY`: Only lighting or color changes.
- `NEAR_DUPLICATE`: Insignificant difference.

- INFOGRAPHIC_DOMINANT: Infographic, annotations, or diagram format.
- TEXT_OR_GRAPHICS_DOMINANT: Dominated by promotional text overlays.
- COLLAGE_OR_COMPOSITE: Collage, split-screen, or comparison.
- INSUFFICIENT_TRANSFORMATION: Negligible change or transformation.

Output Format Output exactly one valid JSON object.

```
{
  "subject_validity": {
    "decision": "PASS" | "FAIL",
    "failure_reason": null | "SUBJECT_ABSENT" | "DIFFERENT_SUBJECT" |
         "CATEGORY_MATCH_ONLY" | "SUBJECT_GRANULARITY_CHANGED" | "
        COMPONENT_PROMOTED_TO_SUBJECT" | "ACCESSORY_SUBSTITUTION" | "
        INSUFFICIENT_REFERENCE_DETAIL" | "UNSEEN_IDENTITY_INFORMATION
        " | "SUBJECT_NOT_IDENTIFIABLE",
  },
  "sdg_appropriateness": {
    "decision": "PASS" | "FAIL",
    "failure_reason": null | "VIEWPOINT_ONLY" | "CROP_OR_ZOOM_ONLY" |
         "DETAIL_CLOSEUP_ONLY" | "COMPONENT_FOCUS_ONLY" | "
        LIGHTING_OR_COLOR_ONLY" | "NEAR_DUPLICATE" | "
        INFOGRAPHIC_DOMINANT" | "TEXT_OR_GRAPHICS_DOMINANT" | "
        COLLAGE_OR_COMPOSITE" | "INSUFFICIENT_TRANSFORMATION",
  },
  "decision": "ACCEPT" | "REJECT"
}
```

Rules:

- If a decision is PASS, failure_reason must be null.
- Output ONLY the JSON object. Do not output markdown, reasoning, or additional text.

Figure 8: Subject-Driven Generation Target Validity Prompt.

B Details of Generating Distortion Controlled SDG Images

Distorting the entire reference would cause the SDG model to render a globally low-quality subject, whereas our objective is for it to reproduce the original high-quality subject with only its fine detail reduced. We therefore confine the distortion to the product's high-frequency, fine-grained regions. Specifically, we distort the full image and then transfer only those regions back onto the clean reference. These regions are localized using Grounding DINO (Liu et al., 2024) and morphological gradients.

However, some artifacts from these localized distortions are unavoidable, as illustrated in Fig. 9. For the controlled set to be meaningful, the three outputs must be realistic SDG results that differ only in fine-grained subject detail, and not in synthetic distortion artifacts. To this end, we distort the reference and regenerate, rather than distorting the subject-driven generated image directly. Because the model re-synthesizes the subject, the resulting output remains a realistic SDG result, free of the artifacts that direct distortion would introduce and differing from a clean generation only in its fine detail. That figure additionally presents examples of the overall distortion, the computed masks and masked distortion results, and the effect of regeneration. We provide a visualization of pairwise evaluation test sets used for Tab. 1 and Tab. 2 in Fig. 10.

Since the None, Mild, and Strong conditions are all SDG outputs, they carry generation-induced distortion. In particular, the None condition is not distortion-free, but is instead the output generated from the undistorted reference. Although the stochasticity inherent to the generative process cannot be eliminated, holding all other factors fixed leaves the source reference as the dominant source of variation among the three outputs.



Figure 9: Visualizations of global, masked distortions and SDG images from masked distorted references.

Table 6: Pairwise Accuracy and AUC_z decomposed by distortion severity, separately reporting original to Mild, Mild to Strong, and original to Strong comparisons.

Category	Metric	original→Mild		Mild→Strong		original→Strong	
		Acc. (%)	AUC_z	Acc. (%)	AUC_z	Acc. (%)	AUC_z
Embedding	CLIP	49.61	4.92	52.41	13.34	52.72	13.15
	DINO	52.72	3.96	63.92	10.14	64.07	10.92
Correspondance	Keypoints	47.36	6.63	84.94	39.48	79.97	38.63
MLLM	DreamBench++	8.40	8.40	13.69	13.69	14.93	14.93
	VIEScore	9.33	5.69	19.13	12.06	20.22	12.59
	MetMag(Ours)	62.67	29.08	84.45	61.67	88.49	68.95

C Detailed Breakdown of Table 1

We first provide a full AUC curves of Tab. 1 in Fig. 11 plotted with Pairwise Accuracies calculated on $k = [0.0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0]$.
We further examine how each metric responds to different distortion severities. As shown in Tab. 6, pairwise accuracy of Ours increases from 62.67% for the subtle Original → Mild comparison to 84.45% for the Mild → Strong comparison and 88.49% for the Original → Strong comparison. This behavior is consistent with the increasing visual difference between the compared outputs, whereas the strong performance on the subtle comparison indicates sensitivity to relatively small changes in fine-grained subject details.
Tab. 6 further reveals a weakness of keypoint matching. Although it performs reasonably well when comparing large differences in detail quality, it struggles to distinguish the subtle differences in the Original → Mild comparison. Moreover, it yields a consistently lower AUC_z than our method across all comparisons, which further confirms the superior detail sensitivity of our method.

D Results for Other SDG Models

Tab. 7 and Tab. 8 show trends consistent with Tab. 1, further demonstrating both the limitations of existing metrics and the advantages of ours.

E Results for Detail Refinement

We further evaluate the effectiveness of Meticulous Magnifier not only for detail distortion but also for refinement. Following the original work, we reimplement FlowFixer (Jun et al., 2026) and apply the controlled distortion experiments from Sec. 3.3 to the refined SDG outputs, which we denote Refined. In total, we conduct 3,855 pairwise evaluations over combinations of the (Refined, None, Mild, Strong) SDG images generated from 643 SDG text prompts.
Tab 9 shows that existing embedding- and MLLM-based metrics cannot even distinguish refined details from distorted ones. Keypoint matching, following FlowFixer, separates these pairs

Table 7: Controlled detail distortion evaluation protocol applied to SDG outputs generated with Flux.1 Kontext[dev], following the same metrics as Table 1.

Metric	Pairwise Accuracy (%)	Tie Rate (%)	AUC_z
CLIP	52.64	0.05	20.68
DINO	53.78	0.00	18.50
DreamBench++	12.89	70.78	12.74
VIEScore	14.80	74.38	12.91
Keypoints	71.84	0.52	37.62
MetMag(Ours)	81.65	3.58	60.82

Table 8: Controlled detail distortion evaluation protocol applied to SDG outputs generated with Qwen Image Edit 2511, following the same metrics as Table 1.

Metric	Pairwise Accuracy (%)	Tie Rate (%)	AUC_z
CLIP	52.80	0.16	16.76
DINO	56.00	0.00	13.96
DreamBench++	14.23	73.65	14.14
VIEScore	19.93	66.77	14.60
Keypoints	67.81	1.29	28.32
MetMag(Ours)	76.31	7.78	48.03

Table 9: Pairwise evaluation results extended to four conditions by including FlowFixer (Jun et al., 2026) refined outputs alongside None, Mild, and Strong.

Metric	Pairwise Accuracy (%)	Tie Rate (%)	AUC_z
CLIP	50.66	0.36	9.65
DINO	57.28	0.23	7.76
DreamBench++	12.53	79.46	12.53
VIEScore	16.06	73.36	10.29
Keypoints	80.34	1.45	36.69
MetMag(Ours)	78.50	6.61	51.79

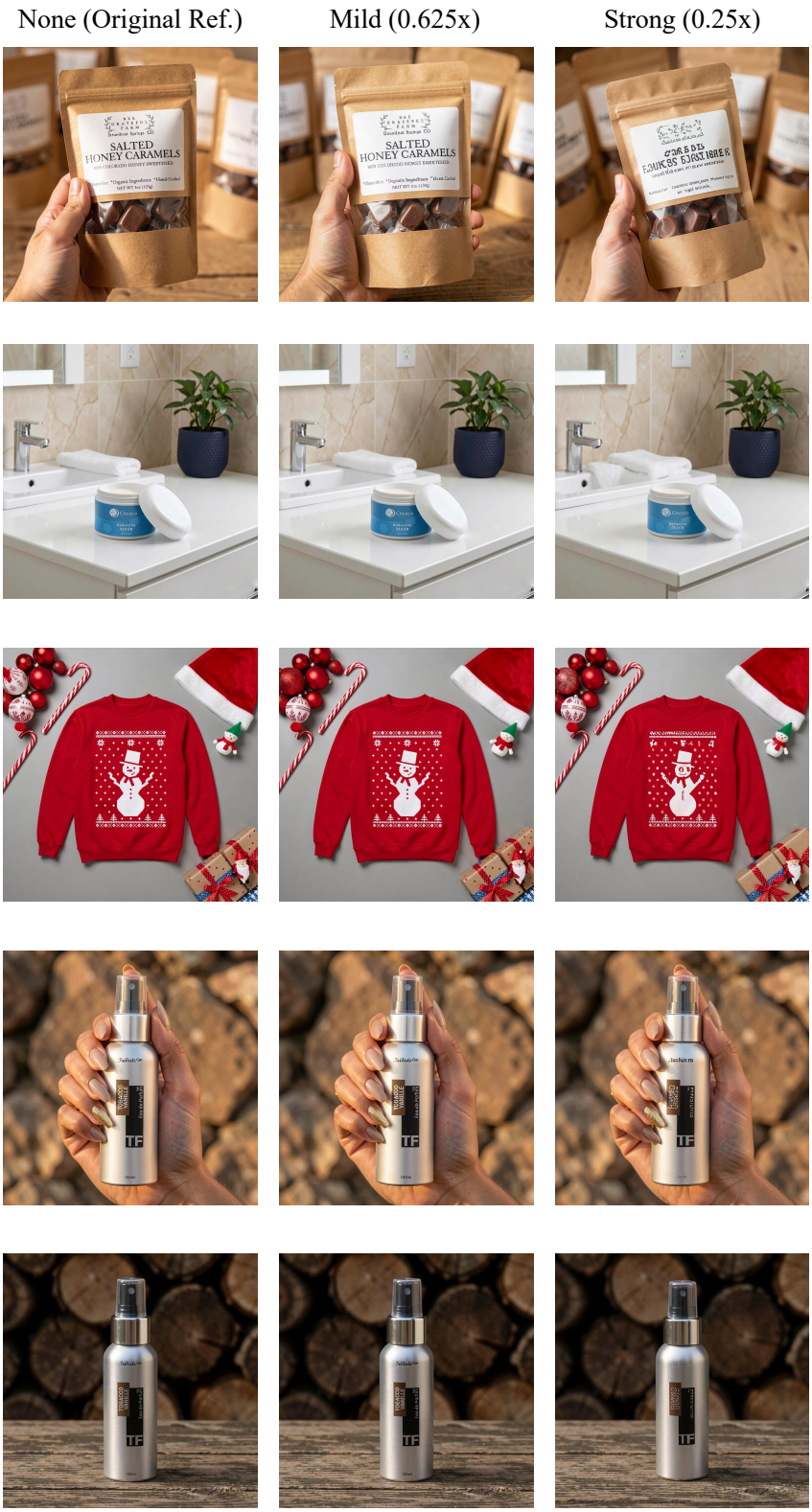

Figure 10: Visualizations of SDG images generated with None, Mild, Strong distorted references. Best viewed in zoomed.

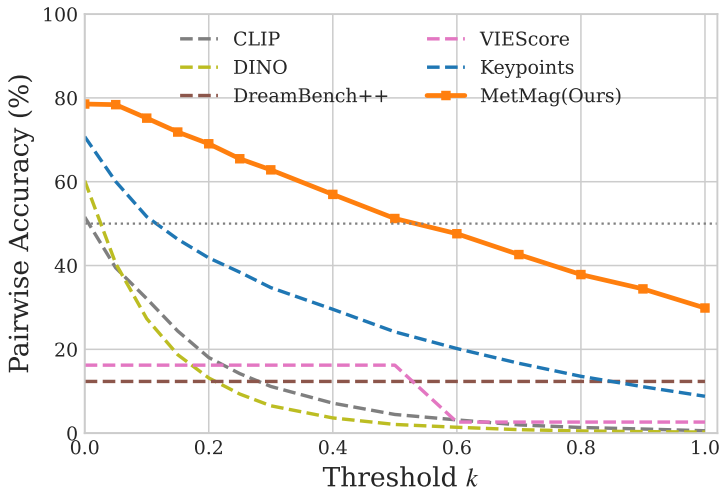

Figure 11: Pairwise Accuracy as a function of the normalized threshold k, for k=[0, 1]. Random Chance (50%) is plotted as a dotted line at the middle.

well—slightly better than ours in Pairwise Accuracy (80.34% vs. 78.50%). However, its lower AUC_z (36.69 vs. 51.79) indicates that it still produces small score margins across different distortion scales.

F Details of Meticulous Magnifier

Task. Given a reference image of a subject, generate a balanced, multi-scale set of visual target labels to be localized and compared for reference-detail preservation.
Each target label identifies a distinct, physically localizable visual region that can be grounded via a bounding box and compared between the reference image and a generated image.
The set must span multiple levels of granularity, covering both:

1. **Coarser Functional & Semantic Regions**: High-level functional modules, structural zones, or grouped attribute assemblies that perform a shared role (e.g., "upper control panel section", "front lacing and tongue assembly", "rear camera housing island").
2. **Salient Components & Features**: Distinct individual parts, structural elements, or prominent identity markings within or across those regions (e.g., "directional navigation pad", "translucent rubber midsole", "embossed gold brand logo").

System Context. The generated titles will be consumed directly as text queries for bounding box (bbox) detection by an open-vocabulary grounding model (e.g., Grounding DINO, OWLv2) and visual evaluator.
For each target title, the system will:

1. Use the title as a detection prompt to predict a bounding box in the reference image.
2. Use the title to predict a bounding box for the corresponding region in the generated image.
3. Compare the visual content within the bounded regions to evaluate fidelity across both macro-structure and localized attributes.

Therefore, each title must function as a standalone, discriminative detection prompt: it must be concise and label-like, yet carry enough descriptive detail to distinguish the intended region or component from other candidates.

Reasoning Process. Before generating target titles, internally perform a top-down visual and functional analysis of the reference subject.

1. **High-Level Decomposition (Functional & Semantic Regions)**:

- Identify the major physical zones, structural sections, and functional modules of the subject.
- Formulate coarse-grained target labels that encompass these semantic assemblies.

2. **Mid-to-Fine Decomposition (Salient Components & Details)**:
   - Within each major zone, identify prominent individual components, mechanical interfaces, and distinctive visual markings (logos, emblems, hardware).
   - Formulate fine-grained target labels for components whose individual preservation is critical to identity.
3. **Balance & Coverage**:
   - Ensure the final set includes both coarse functional areas (capturing overall layout and relationships) and specific key components (capturing fine details).
   - Prune redundant or trivial micro-elements that do not contribute meaningfully to visual evaluation.

Do not output the inspection process. Output only the final JSON array.

Requirements. Each target title must satisfy the following requirements.

1. **Discriminative & Specific**
   - Provide sufficient visual specificity to distinguish the target region or component from all other candidates in the image.
   - Include discriminating modifiers such as location, color, material, or functional role (e.g., use "upper front control cluster" or "lower rubberized outsole" rather than ambiguous general terms).
   - For components with attached text or icons, incorporate the marking as a descriptive modifier of the component rather than creating an isolated box for the text.
2. **Concise Noun Phrase (Detection Label Format)**
   - Format each title strictly as a compact noun phrase suitable for open-vocabulary object detectors.
   - Keep titles brief (typically 2 to 6 words).
   - Do not write full sentences, explanatory clauses, or prose descriptions.
   - Do not use slashes ("/") or disjunctions ("or") to combine disparate alternatives into one label.
   - Do not include meta-phrases such as "a photo of", "visual attribute of", or "bounding box for".
3. **Bbox-Groundable Across Scales**
   - Every target, whether a coarse functional zone or an individual component, must correspond to a coherent visual area encloseable by a single 2D bounding box.
   - For coarse regions, the bounding box frames the entire functional module or structural section as a cohesive whole.
   - For individual components, the bounding box frames the specific component.
   - In all cases, the box must cleanly frame the target without spanning across unrelated distant areas or large empty gaps.
4. **Multi-Scale Representation & Cohesion**
   - Treat multi-action controls, contiguous rows/grids of repeated elements, or shared functional strips as single cohesive targets.
   - Include both broad functional assemblies and key constituent components, but avoid extreme over-fragmentation into trivial sub-features.
   - Avoid duplicate titles that target the identical bounding area under different names.
5. **Observable**
   - Describe only visual elements directly visible in the reference image.
   - Do not infer unseen internal parts, invisible electronics, or unobservable brand lore.

Generation Guidelines.

- Formulate every title following standard detection prompt syntax: `[optional position/material/functional modifier] + [region or component name]`.
  - Regional / functional examples: “upper media playback control strip”, “front lacing system”, “rear dual camera housing island”, “padded ankle collar assembly”.
  - Component-level examples: “directional navigation pad”, “gold engraved buckle”, “black rubberized outsole tread”, “embossed circular logo”.
  - Poor examples: “buttons” (too generic), “the panel that controls the volume and channels” (sentence-like prose), “premium quality finish” (not bbox-groundable).
- Maintain a balanced distribution across scales: approximately one-third coarse functional/semantic regions and two-thirds salient component-level targets.
- Exclude non-intrinsic scene elements (background, shadow, lighting reflections, display stand).

OUTPUT FORMAT. Return your response as a valid JSON array. Each element must follow the structure below.

```
[
  {
    "title": "<concise, discriminative bounding box detection label>"
  }
]
```

Figure 12: Multi-Scale Visual Target Label Generation Prompt

F.1 SALIENT TARGET LIST GENERATION

We instruct Qwen3-VL-Instruct-32B to represent each subject at multiple levels of granularity. Fine-grained targets capture distinctive components, markings, and localized details, such as a “red power button with LED indicator” or an “embossed gold brand logo.” However, focusing only on fine-grained details may overlook relationships among individual components or broader structural configurations. We therefore additionally include coarser functional and semantic regions, such as an ”upper control panel section” or a ”rear camera housing island.” This multi-scale representation allows the evaluation to cover both broader subject structure and distinctive local details. Fig. 12 shows the text prompt used for the list generation, and Fig. 13 visualizes examples of generated list. Each target is represented by a concise textual title that describes the visual region itself. Because these titles are subsequently used as open-vocabulary queries for grounding, we constrain them to be concise and discriminative noun phrases, typically consisting of 2–6 words. Positional, material, color, or functional modifiers may be included when necessary to uniquely identify the intended region. Each target should correspond to an observable and spatially coherent region that can be localized with a bounding box, while abstract properties or regions that cannot be visually isolated are excluded.

The target titles serve only as localization queries and are not provided to the final evaluator. Instead, the evaluator receives only the corresponding reference and generated crops. Thus, the target-generation stage determines the visual criteria and their locations, while the final MLLM judge evaluates the visual correspondence between the localized regions without being given target-specific textual descriptions.

F.2 OPEN-VOCABULARY GROUNDING AND CROP EXTRACTION

For consistency, the bounding box in the reference image is detected once and reused across all subsequent evaluations of the same subject, whereas each target is localized independently in every generated image.

To provide sufficient local context and mitigate minor localization errors, we expand each detected bounding box by 15% of its width and height on each side before cropping. When the shorter side of a crop is smaller than 224 pixels, we resize the crop so that its shorter side becomes 224 pixels while preserving its aspect ratio.

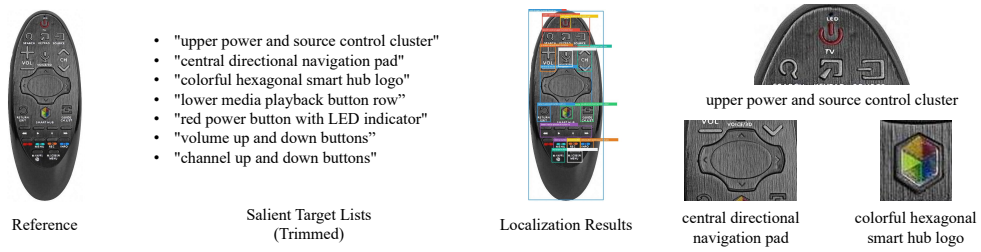

Figure 13: Visualizations of salient target list and corresponding bbox detection results and crops.

Table 10: Ablation on the minimum crop side length before upsampling, reporting overall Pairwise Accuracy(Acc.) and AUC_z with a breakdown by severity. **Bold** indicates the best (rank 1) value and underline indicates the second-best (rank 2) value in each metric column.

Crop	3-step Summary			original→Mild		Mild→Strong		original→Strong	
	Acc. (%)	Tie (%)	AUC_z	Acc. (%)	AUC_z	Acc. (%)	AUC_z	Acc. (%)	AUC_z
56px	77.03	7.57	51.21	59.72	29.13	84.60	58.98	86.78	65.52
112px	77.92	**6.07**	51.15	60.19	29.11	**86.63**	**59.04**	86.94	65.30
224px	**79.21**	6.17	51.28	63.14	29.06	86.00	58.66	88.49	66.13
448px	78.64	7.52	**51.54**	**63.45**	**29.70**	83.67	58.60	**88.80**	**66.33**

To assess the effectiveness of crop upscaling, we ablate the minimum crop side length applied before upsampling small crops, varying it from 56 to 448 pixels. Across the full evaluation set, overall Pairwise Accuracy and AUC_z remain relatively stable across crop resolutions, suggesting that our method is not highly sensitive to this hyperparameter in aggregate. Decomposing performance by distortion severity, however, reveals a more informative trend. On the subtlest comparison (Original→Mild), strict concordance increases monotonically with crop resolution, from 59.72% at 56px to 60.19%, 63.14%, and 63.45% at 112px, 224px, and 448px, respectively. The same pattern holds for the Original→Strong comparison (86.78%, 86.94%, 88.49%, and 88.80%). This indicates that although low-resolution crops suffice to distinguish coarse-grained distortions, higher crop resolution specifically improves sensitivity to subtle, fine-grained differences. This motivates our choice of 224px, which captures most of this benefit without unnecessarily inflating the input size.
We instruct Qwen3-VL-Instruct-32B to conduct a multi-object detection by providing the model of an image and corresponding entire Salient Target List. Figure 14 shows the text prompt used for this process along with a qualitative example.

F.3 Crop-Level MLLM Judging

We provide the MLLM a reasoning scaffold that instructs the model to reason about the fidelity of visual properties in the crop that are relevant to the target. We explicitly exemplify criteria to consider, such as correctness, completeness, readability, geometry, structural consistency, detailed appearance, and relative placement. Fig. 15 shows the text prompt used for crop-level judgment along with an example answer.

G Performance of Open-Vocabulary Bbox Detection

G.1 Detection Accuracy

Since our pipeline relies on part-level boxes predicted by Qwen-32B (Bai et al., 2025), we verify their quality through a human study. As shown in Tab. 11, Three raters independently inspected 100 randomly sampled product images (1,431 boxes across 23 categories) and judged whether every box tightly covers its labeled part. Under a strict criterion that counts an image as correct only when all three raters agree, 89% of images are fully correct, and none were judged to localize the wrong object. Most of the remaining errors are borderline cases flagged by a single rater, and inter-rater agreement is high (Gwet's AC1 = 0.92).

Table 11: Human evaluation of part bounding boxes. An image is counted as *OK* only if all three raters judge it correct. *# Raters flagging*: number of raters who marked the image as an error. Brackets denote 95% Wilson intervals.

	Images		# Raters flagging			Boxes	
Verdict	Count	Ratio (%)	1	2	3	Total	Per image
OK	**89**	**89.0** [81.4, 93.7]	–	–	–	1,264	14.2
Error	11	11.0 [6.3, 18.6]	9	1	1	167	15.2
Truncated	3	3.0	2	0	1	44	14.7
Missing	6	6.0	6	0	0	90	15.0
Too wide	2	2.0	1	1	0	33	16.5
Other	0	0.0	0	0	0	0	–
Total	100	100.0	9	1	1	1,431	14.3

G.2 Detection Failure Rate

On average, each reference subject yields 14.16 salient targets, with an average of 0.43 missing detections per reference–original SDG pair, corresponding to a target-level detection failure rate of 2.7%.

Locate every instance that belongs to the following categories: *[Salient Target List]*. Report bbox coordinates in JSON format.

```
<output_format>
```json
[
  {{"bbox": [y1 x1 y2 x2], "description": "{sample_cat}: detailed
      description of visual appearance..."}},
  ...
]
```
</output_format>
```

Figure 14: Open Vocabulary BBox Detection Prompt

You are an expert evaluator of subject identity preservation in commercial image generation. Your task is to evaluate how faithfully the generated image preserves the identity attributes described by the provided evaluation rubrics.

Inputs.

- Reference Image (Image 1)
- Generated Image (Image 2)

Treat each rubric item as an independent visual verification task.
Do not perform a global or holistic comparison between the entire images; evaluate each rubric item independently.
For each individual rubric item, follow Steps 1 through 5 sequentially and internally. Do not include your internal reasoning in the final output.
Use the reference image (Image 1) solely to locate and understand the visual attribute described in the rubric item. Assess the preservation and visibility of that attribute using only the generated image (Image 2). The status must be based strictly on direct observations in the generated image (Image 2). Do not infer or assume any missing details based on the reference image, prior knowledge, or brand familiarity. If an attribute cannot be observed with confidence in the generated image (Image 2), its status must be set to "Missing".

Step 1. Verify the Identity Attribute. Assess the identity attribute specified in the generated image (Image 2).

When appropriate, consider
- correctness
- completeness
- readability
- geometry
- structural consistency
- detailed appearance
- relative placement

Do not infer missing information. Do not just rely on overall semantic similarity. Errors or discrepancies in unrelated attributes must not affect the evaluation.

STEP 2. DETERMINE THE STATUS. Assign a specific preservation status based on the preservation quality of the specified identity attribute (considering correctness, completeness, readability, geometry, structural consistency, detailed appearance, and relative placement).
- **Very Poor**: The identity attribute is incorrect, represents a completely different identity/value, or is missing entirely (e.g., incorrect product name, wrong logo, incorrect printed number/text).
- **Poor**: Minimal preservation. The attribute is recognizable but suffers from severe degradation, distortion, critical structural/geometric errors, or is largely unreadable.
- **Fair**: Moderate preservation. The attribute shows clear resemblance to the reference with a few notable discrepancies (e.g., partial text readability issues, incomplete graphics, or moderate geometric variance).
- **Good**: Strong preservation. The attribute closely matches the reference across appearance, geometry, and readability, but minor discrepancies or slight imperfections exist.
- **Excellent**: Perfect preservation. The specified identity attribute is faithfully and completely preserved across all detailed criteria with no noticeable discrepancies.

STEP 3. PRODUCE THE EVALUATION. For each rubric item, produce
- title
- status
- visual_evidence

The `visual_evidence` field must identify the observed details and explain how they justify the selected status.
Keep the explanation of visual evidence concise (1–2 sentences).

OUTPUT FORMAT. Return a JSON array. Each JSON object must contain exactly the following keys.

```
{
   "title":"...",
   "status":"Very Poor" | "Poor" | "Fair" | "Good" | "Excellent",
   "visual_evidence":"..."
}
```

RULES.
- The title must exactly match the rubric title.
- Produce exactly one evaluation for every rubric.
- Evaluate every rubric independently.
- Do not summarize the overall image.
- Do not aggregate scores.
- Do not output your reasoning.

Figure 15: Crop-level Subject Identity Preservation Evaluation Prompt

H REFERENCE/GENERATED IMAGE ORDER BIAS

We additionally verify robustness to the order in which the reference and generated crops are presented to the MLLM judge by reversing their order (generated-then-reference instead of reference-then-generated).

Table 12: Effect of reference/generated presentation order.

Order	Strict Conc. (%)	Tie (%)	AUC_z
Reference → Generated (Ours)	78.54	6.69	53.23
Generated → Reference (Reversed)	75.53	8.14	51.04

Strict Pairwise Accuracy decreases slightly from 78.54% to 75.53%, and AUC_z decreases from 53.23 to 51.04, indicating a modest sensitivity to presentation order while remaining substantially above all baseline metrics evaluated in Sec. 3.3.