DOMAIN-EDIT: A BENCHMARK FOR DOMAIN-UNBIASED EVALUATION OF GENERATIVE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

As text-to-image models advance, many of the prominent compositional generation benchmarks are saturating. Yet, modern generation models have been shown to be limited in their generation diversity. This makes it difficult to assert whether the generative models have developed general abilities, or their abilities have become specialized. We observe that these models are also biased in domain (like style, layout, object properties), a phenomenon we dub *generator-domain bias*. Existing compositional benchmarks do not control for *generator-domain bias*, so models are free to fall back on their preferred domain. We argue that editing tasks serve an appropriate medium for studying this: unlike pure generation, image editing requires the model to compose within a source image it did not choose. To this end, we introduce **DOMAIN-EDIT**, a benchmark that pairs compositional editing instructions with source images generated across diverse styles. We evaluate eight open-source models on three subsets of increasing visual diversity, spanning plain textures and five styles. We find that every model falls below its own generation accuracy once the domain is fixed, that performance varies substantially with the source style, and that models often regenerate the background rather than edit it. Based on these findings, we argue that input-domain diversity should be a core dimension of compositional evaluation.

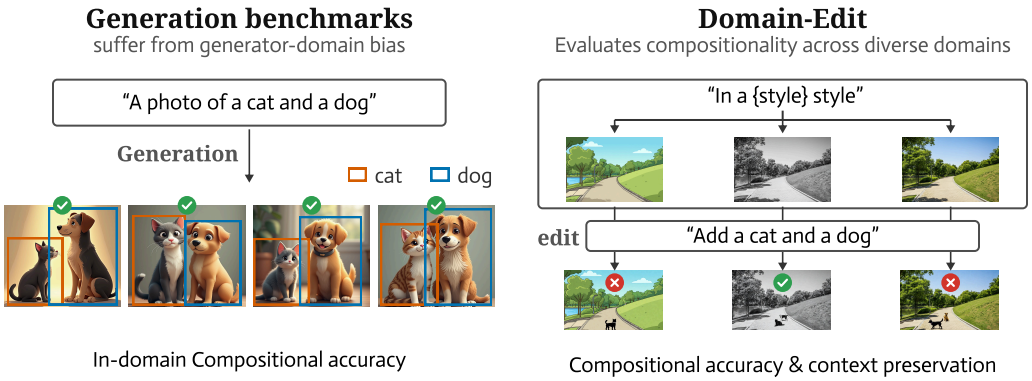


Figure 1: **Generation benchmarks vs. DOMAIN-EDIT.** *Left:* In generation benchmarks, the model can chooe its own domain. For example, given "A photo of a cat and a dog", all generations often look similar, so a high benchmark score reflects performance only within the model's preferred domain. *Right:* DOMAIN-EDIT fixes the domain through a source image rendered in different styles and asks the model to add the same objects.

1 INTRODUCTION

As text-to-image (T2I) generation models advance rapidly, they demonstrate an impressive ability to produce high-quality images. A number of benchmarks have been proposed to measure these abilities Ghosh et al. (2023); Huang et al. (2023). Recent models score highly on them, with accuracies approaching 100%, giving an impression that contemporary models have developed compositional understanding. Despite this strong performance, however, these models have been shown to exhibit limited output diversity Adamkiewicz et al. (2026). Importantly, they are also biased in *domain*, aspects like layout, style, and object properties, a phenomenon we dub *generator-domain bias* (Figure 1). This raises the question: are models performing well because they have developed an understanding of visual concepts, or is their understanding shaped by generator-domain bias?

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.

To examine *generator-domain bias*, we break a domain down into two components: layout and style. We find that current models are strongly biased along both. Given the same prompt, generations across random seeds place and size objects in nearly identical configurations: the bounding boxes of the same object overlap with a high average IoU across seeds. Likewise, nearly all generations share a single style, regardless of the seed. We do not claim that such consistency is undesirable in itself. Rather, it means that benchmark scores reflect performance within each model's preferred domain, leaving open whether they hold outside it.

Studying *generator-domain bias* requires testing models beyond their preferred domains. A natural test is to check whether the model can produce specified images in a different domain. In standard generation, we sample from $p(x \mid y)$, where y is some text prompt. But y underspecifies everything outside the compositional target: style, background, layout, object appearance. (e.g., specifying a watercolor style leaves many interpretations for the model to realize) The model fills these in from its own prior, which is exactly what *generator-domain bias* affects. In this work, we aim to control the domain in the compositional image generation setting.

Concretely, we argue this problem can be studeid through image editing. Given a context image x', we propose to evaluate $p(x \mid y, x')$ under the typical prompts y used in standard compositional generation benchmarks. Here, x' acts as a hyper-specific prompt for everything y cannot express, while y isolates the compositional target. The model cannot fall back on its preferred layout or style. It must place and modify objects within a context it did not choose, revisiting saturated compositional benchmarks under conditions the model no longer controls.

Layout is controlled implicitly through the objects present in the source image, while style is varied explicitly. We organize our benchmark into three subsets whose source images increase in visual diversity. The Texture subset consists of plain textures without objects and serves as a layout-free baseline. The Landscape subset consists of background-only scenes rendered in different styles. For both, we follow the GenEval protocol but deliver each compositional prompt as an edit on the source image, and test whether compositional ability is preserved uner domains the model did not choose. The Scene subset takes object-centric images and editing instructions from GEdit-Bench Liu et al. (2025) and re-renders the source images across styles, testing whether editing ability itself is robust to context. Across all three subsets, performance depends on the style of the source image, even though the instruction is unchanged.

Our main contributions are summarized as follows:

- We highlight the limitation of pure generation and realistic -only editing benchmarks in evaluating compositionality, formally defining the concept of *generator-domain bias* in T2I editing models.
- We introduce DOMAIN-EDIT, a benchmark of three subsets (Texture, Landscape, and Scene) that evaluates compositional editing across source styles, measuring how much performance drops relative to generation.
- We provide a comprehensive empirical analysis using diverse open-source models, revealing that current models exhibit severe vulnerability to input style variations when executing identical editing instructions.

2 RELATED WORK

Generative AI models and diversity. Recent diffusion Rombach et al. (2022); Podell et al. (2024), flow-matching models Esser et al. (2024), and unified models Deng et al. (2025); Xie et al. (2025) have made rapid progress in text-to-image (T2I) generation, producing complex, high-quality scenes from text prompts. Building on these models, instruction-guided editing models Brooks et al. (2023); Liu et al. (2025); Labs et al. (2025) edit an image according to a text instruction while preserving the regions the instruction does not target. Despite their improving visual fidelity and semantic alignment, recent generative models have been shown to produce images of limited diversity Adamkiewicz et al. (2026). Several works further point to domain bias in generative models. Jeong et al. (2025) observe that recent models are biased toward the domains they generate, and Fathi et al. (2026) study content and style biases in diffusion classifiers. For RL-tuned models, DivGenBench Chen et al. (2026) shows a trade-off between human preference and generative diversity across several dimensions, including layout and style. We build on these observations and test whether such biases affect compositional ability when the domain is fixed by a source image.

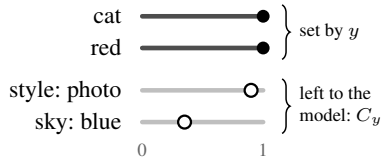
Generation and editing benchmarks. Numerous benchmarks, such as GenEval Ghosh et al. (2023), T2I-CompBench Huang et al. (2023) and DPG-Bench Hu et al. (2024), have been introduced for T2I generative models. They assess whether a model adheres to prompts involving attribute binding, object counting, and spatial relationships. However, state-of-the-art models Diao et al. (2026); Wu et al. (2025b) are rapidly approaching saturation on these leaderboards, achieving near-perfect scores. More recent benchmarks such as GenEval2 Kamath et al. (2025) and T2I-CompBench++ Huang et al. (2025) respond by making the prompts harder: more objects, finer-grained attributes and their combinations, and longer prompts. Editing benchmarks such as GEdit-Bench Liu et al. (2025), EditVal Basu et al. (2023), and Emu Edit Sheynin et al. (2024) provide one source image that fixes the visual context, so the model cannot freely choose it. They check both whether the requested change was applied and whether the rest of the image was preserved, which requires the model to understand the given context.

Style in generative models. Work on style in T2I generation has mostly focused on style transfer and on separating style from content. Large-scale datasets such as OmniStyle-1M Wang et al. (2025), Style30K Li et al. (2024), StyleExpert-500K Zhu et al. (2026), and MegaStyle Gao et al. (2026) categorize stylized images or decouple style from semantic content. For evaluation, HPSv2 Wu et al. (2023) scores generations across styles such as animation and concept art, and UnlearnCanvas Zhang et al. (2024) measures how well unlearning methods remove specific styles from a model. Instead, we use stylized images, such as sketches and cartoons, as fixed source contexts to test whether compositional ability holds across domains.

3 Generative models exhibit Generator-Domain bias

In this section, we illustrate the pervasiveness of domain bias in generative models. We hypothesize that *generator-domain bias* is a problem that arises in models that aim to perform well on T2I generation benchmarks due to saturation of style. This manifests in limited diversity (Jeong et al., 2025), model-specific stylistic choices (Adamkiewicz et al., 2026), and content/style biases Fathi et al. (2026). To make the discussion concrete, we define the notion of a domain.

3.1 Defining generator-domain bias



Generator-domain bias is the tendency of a generative model to favour particular visual realizations of concepts left unspecified by the prompt. Intuitively (see fig. on the right), we can think of the image generation process as a very large mixing board of levers over a set of concepts and their values, each taking a value from 0 to 1, expressing how much a certain concept value needs to be realized (say the presence of a cat, its size, presence of other objects, style of each object, or pixel or whatever else specification one could think of). A text prompt specifies just a number of these, leaving all others free to be realized. The generative model itself needs to pull the remaining levers as appropriate. The settings of these free levers that a model actually produces are what we refer to as its domain, and the tendency of a model to pull them to the same settings in a systematic way is what we refer to as bias. We now specify these terms more formally, starting from the definition of domain, where we define it with respect to concepts similar to Uselis et al. (2026).

Definition 1 (Domain of a generative model $p(\cdot \mid y)$)**.** *For a given text prompt y, let C_y denote the set of concepts that are not specified in y, where each concept $c \in C_y$ can take values in a set V_c compatible with y. For an image X, let $c(X) \in V_c$ be the value that concept c takes. The domain of a generative model $p(\cdot \mid y)$ on a concept $c \in C_y$ is the distribution P_c^y of $c(X)$, where $X \sim p(\cdot \mid y)$.*

For example, for a prompt like $y :=$"a red cat", its set of concepts that are not specified (thus included in C_y) is all kinds of things: presence of a helicopter, style of the image, pose of the cat, etc. For the style, typical values may be $V_{\text{style}} = \{\text{photo}, \text{sketch}, \dots\}$, while the position of the cat may take continuous values, $V_{\text{position}} = [0, 1]^4$, specifying bounding boxes. The domain of a model then describes how often it produces each value, e.g., a model that only ever generates photos of the cat in the centre of the image, mostly sitting and sometimes lying, puts all its mass on photo for the style and on boxes around the centre for the position, and most of it on sitting for the pose, even though y allows many more values (Figure 2).

Definition 2 (Generator-domain bias)**.** *Let M be a measure of how concentrated a density/distribution over V_c is. The generator-domain bias of a model $p(\cdot \mid y)$ on a concept $c \in C_y$, as measured by M, is $M(P_c^y)$, where P_c^y is the domain of p on c (Definition 1).*

For example, M can be the probability of the most frequent value, the (negative) entropy, or the expected similarity between two independent generations. In Section 3.2, we take c to be the layout of objects and M the average IoU of their bounding boxes across seeds.

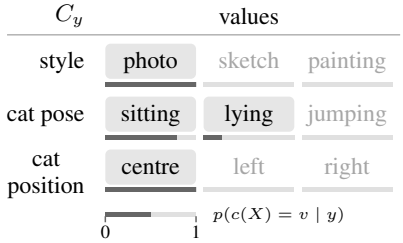


Figure 2: **Domain of an examplar model $p(\cdot \mid y)$ for $y =$ "a red cat"**. Rows are concepts in C_y, with some of their values, all compatible with y. Bars show the domain P_c^y: how often the model produces each value v.

3.2 Evidence of generator-domain bias

Here, we investigate to which extent *generator-domain bias* is present in the current leading models. We adopt Definition 2 to measure *generator-domain bias*, which is defined over a measure M over the distribution of concept values. We mostly focus on three related concepts and their measures: object layout, object size, and style. Prior work typically details bias stylistically over metrics that rely on embeddings (Naeem et al., 2020; Adamkiewicz et al., 2026), where these concepts are entangled, we instead pin down the bias at concept level.

Setup. We consider GenEval (Ghosh et al., 2023) as our testing ground, consisting of *single object*, *two objects*, *color*, *color attribution*, *counting* and *position* categories, from which we take 60 prompts (10 per category). For each prompt y, we generate $S = 5$ images with varying seeds s, and use targeted grounding with Qwen3.5-9B (Qwen Team, 2026) to extract the bounding box $b_{s,o}$ of each object o mentioned in y. We keep an object only if exactly one box is found for it, except for counting prompts, where we keep all instances. We then instantiate Definition 2 with two concepts in C_y for each object o: its *layout*, whose value is the bounding box $b_{s,o} \in [0,1]^4$ capturing both its position and extent, and its *size*, the area $a_{s,o}$ of that box, neither of which is fully specified by the prompts. As measures of concentration M, we compare the object across pairs of seeds,

$$M_{\text{layout}} := \frac{1}{\binom{S}{2}} \sum_{s<s'} \text{IoU}\big(b_{s,o}, b_{s',o}\big), \qquad M_{\text{size}} := \frac{1}{\binom{S}{2}} \sum_{s<s'} \frac{\min(a_{s,o}, a_{s',o})}{\max(a_{s,o}, a_{s',o})}, \tag{1}$$

i.e., the average IoU and the average area ratio of the object's boxes over all pairs of seeds. M_{size} is 1 if the object has the same size in both generations and 0.5 if it is twice as large in one of them. Higher values mean that the model keeps placing the object in the same layout, or at the same size, respectively. We average both over the objects of a prompt and then over prompts; for counting prompts, we match the instances of two seeds one-to-one with the Hungarian algorithm before comparing them. For the *style* of the image, whose value is its primary style (e.g. photo, 3D render, cartoon) as labelled by a Qwen3.5 9B VLM (Qwen Team, 2026), we use the share of images in the most frequent style, $M_{\text{style}} := \max_v \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}[\text{style}(X_i) = v]$ over the N images of a model, computed on prompts that do not specify a style. Refer to Section 5 for experimental settings.

Results: models generate layouts and sizes of their own choosing. Figure 3 shows that all models strongly concentrate both concepts, although neither is specified by the prompts. Across models, M_{layout} ranges from 0.53 (OmniGen2) to 0.72 (FLUX.2), and M_{size} is even higher, from 0.71 to 0.81 (Figure 3a); BAGEL behaves similarly on its own prompts (0.50 and 0.68). Style is concentrated even more strongly: without a style in the prompt, M_{style} ranges from 0.86 (OmniGen2) to 1.0 (Qwen-Image), with photo as the dominant style for every model; BAGEL, the least concentrated at 0.72, departs mostly towards 3D renders (Figure 3b).The concentration is strongest for single objects and weakest for two objects (Figure 3c), where the prompt constrains the scene more, yet it stays high throughout. Even a typical prompt such as "a photo of three birds" ($M_{\text{layout}} = 0.71$, close to FLUX.2's average) yields nearly the same arrangement in every seed: three birds of the same size, side by side, at the same height (Figure 3d). The same holds across prompts: box centres pile up in a few spots, e.g. a single spot for single objects and four fixed spots for position prompts in FLUX.2, and while OmniGen2 spreads them more, it favours the same regions (Figure 3e).

In sum, a generation benchmark such as GenEval only evaluates a model on the layouts the model itself prefers; this does not tell us whether its compositional abilities hold beyond them. This is what image editing lets us test: by providing the source image, we fix the canvas ourselves.

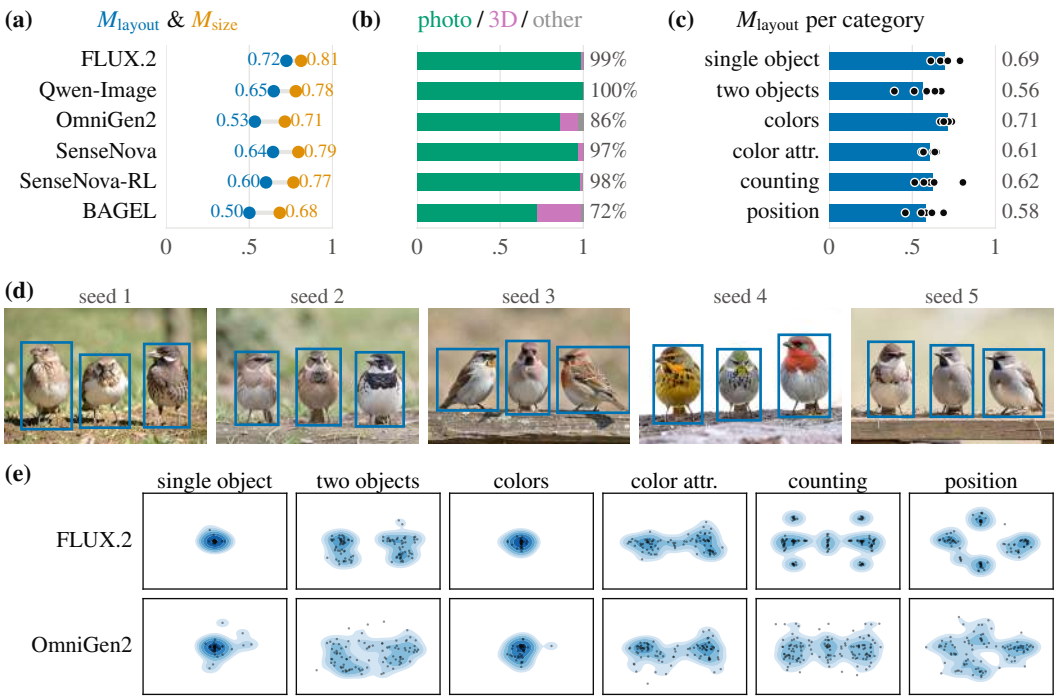


Figure 3: **Layout and size concentration on GenEval across seeds. (a)** M_{layout} (mean IoU of object boxes across seeds) and M_{size} per model, averaged over prompts. **(b)** Distribution of image styles per model for prompts that do not specify a style; the number is M_{style}, the share of photos. **(c)** M_{layout} per category: bars are the mean over models, dots the individual models. **(d)** All five seeds of FLUX.2 for $y :=$ "a photo of three birds" ($M_{\text{layout}} = 0.71$), with their detected boxes: the model produces nearly the same arrangement every time. **(e)** Box centres and their density, for FLUX.2 and OmniGen2 models. Although prompts leave layout and size unspecified, all models concentrate them strongly.

4 DOMAIN-EDIT BENCHMARK

In this section we overview our approach of studying *generator-domain bias* which we established in Section 3. Figure 4 gives an overview of the benchmark and its evaluation.

Preliminaries. In text-to-image generation, a model generates an image $x \sim p(x \mid y)$ from a text prompt y and is free to choose the style, layout, and background. In image editing, the model additionally receives a source image x' and samples $x \sim p(x \mid y, x')$, so it must realize y within the context given by x'.

Data curation. DOMAIN-EDIT evaluates whether a model's compositional ability holds when the domain is not specified by a model; instead, the domain is largely specified by images. All subsets follow the same principle: the instruction y is held fixed while the source image x' is varied across domains. We consider two variants (Figure 5, Table 5). *Generation-derived* subsets (Figure 4a) start from a generation benchmark and turn its prompts into edits, so that editing performance can be compared directly with generation from the same prompts. *Editing-derived* subsets (Figure 4b) start from an existing editing benchmark. Such benchmarks already fix the domain through the source image, but their sources are predominantly photographs. We therefore re-render the same sources in several styles, keeping the instructions unchanged, to test whether editing ability measured on photos holds in other domains. See Section A for details.

- *From generation.* We build two generation-derived subsets on GenEval Ghosh et al. (2023): *Texture* and *Landscape*. Each GenEval prompt (single object, two objects, counting, colors, position, or color attribution) is rephrased as an edit, e.g., "add a dog to the image". These subsets differ in how complex the source image is. *Texture* (Figure 4a-i) uses plain textures and single-color backgrounds, with no objects or layout. *Landscape* (Figure 4a-ii) uses background-only scenes, each shown in five styles: photorealistic, cartoon, painting, grayscale, and sketch.
- *From editing.* We build one editing-derived subset on GEdit-Bench Liu et al. (2025): *Scene*. Its source images show objects in everyday scenes, and about 90% of them are photorealistic. We

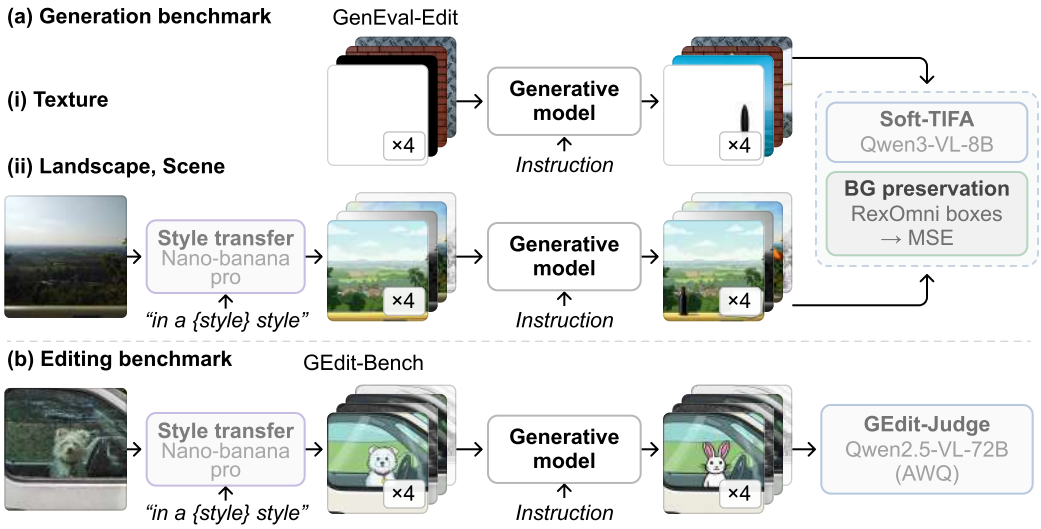


Figure 4: **Overview of DOMAIN-EDIT. (a)** Generation-derived subsets: GenEval prompts are posed as edits on (i) plain textures and (ii) background-only landscapes, which are style-transferred into five styles. **(b)** Editing-derived subset: GEdit-Bench sources are style-transferred into five styles, with the original instructions kept unchanged. In all subsets, the instruction is fixed and only the source image varies. Edits are scored with Soft-TIFA and background MSE (a) or the GEdit-Bench judge (b).

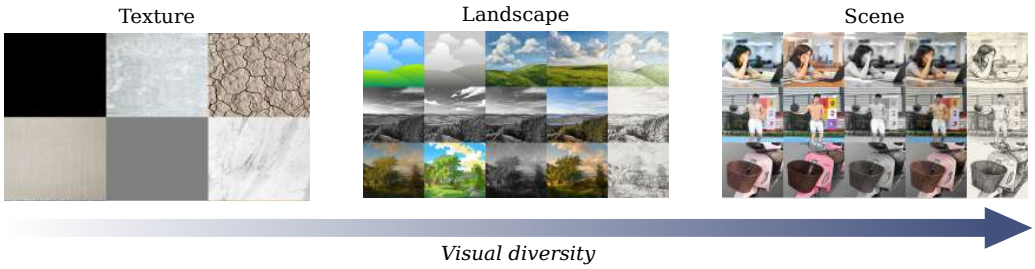


Figure 5: **Source images of the three subsets**, from least to most complex. **Texture:** plain textures and single-color backgrounds. **Landscape:** background-only scenes (rows) in five styles (columns). **Scene:** GEdit-Bench images in five styles.

keep photorealistic images and re-render each one in the four other styles: cartoon, painting, grayscale, and sketch. Instructions are unchanged, so only the source style varies.

Together, the subsets cover sources ranging from plain backgrounds with no content to dense real-world scenes. The generation-derived subsets measure how much a model loses relative to its own generation, and the editing-derived subset measures how much existing editing scores depend on photographic sources. Details on source collection and prompts are given in Section A.

Evaluation. We score each edit on whether it follows the instruction (semantic alignment) and whether it leaves the rest of the source intact (source preservation), following existing benchmarks (Figure 4, right). For Landscape, we additionally introduce a context score. Table 5 lists the metrics and judges used for each subset. Implementation details are given in Section A.

Semantic alignment. For Texture and Landscape, we use Soft-TIFA (Kamath et al., 2025), which splits each prompt into atomic questions and answers each with a VLM (Qwen3-VL-8B). Unlike the object detector of GenEval, which is trained on photos, a VLM judge also works on stylized images. We aggregate the answers with a geometric mean and denote as Acc. For Scene, we follow GEdit-Bench and use the semantic consistency (SC) score of VIEScore.

Source preservation. A high semantic alignment score does not by itself mean the edit succeeded: a model can add the correct objects while regenerating the background in its preferred style, which defeats the purpose of fixing the domain. For Texture and Landscape, we therefore measure how much the background changes: we detect the added objects with Rex-Omni (Jiang et al., 2026), mask them out, and compute the MSE between the output and the source on the remaining background (lower is better). For Scene, we follow GEdit-Bench and use the perceptual score VIEScore.

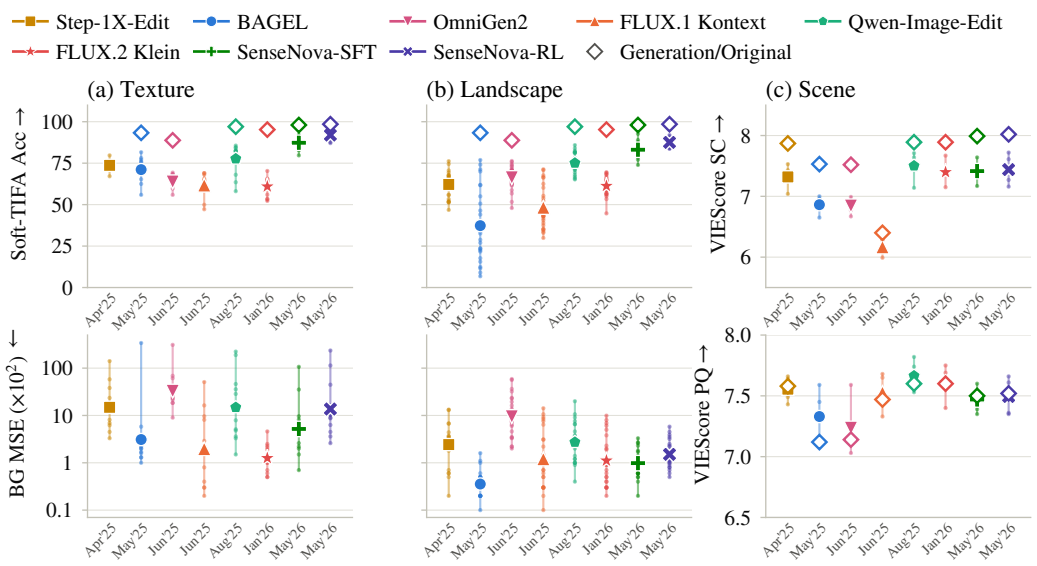


Figure 6: **Editing robustness across input domains.** Each column is one editing model. Small dots are individual input conditions (12 textures in (a), 5 styles in (b), 5 styles in (c)); the vertical line spans their min–max, and the large marker shows the geometric mean over conditions. The diamond marker shows the same model's generation-only accuracy except Qwen-Image, i.e., from the prompt without a source image. **(a)** Texture and **(b)** Landscape: Soft-TIFA accuracy (top) measures whether the edit was applied, and background MSE (bottom, log scale) measures how much of the untouched region changed. **(c)** Scene: VIEScore semantic consistency (SC, top) and perceptual quality (PQ, bottom), averaged over the five core tasks. Larger dot spread indicates stronger dependence on the input domain.

Context score. We observe that edits often contain the correct objects, yet these are poorly integrated into the scene, e.g., objects floating in mid-air or with the wrong scale or lighting. Since neither semantic alignment nor source preservation captures this, we introduce a context score for Landscape (rated by a VLM; details in Section A.4). It does not apply to Texture, whose plain backgrounds offer no scene to fit, nor to Scene, which we evaluate with the judges of GEdit-Bench.

5 Experiments

We use Domain-Edit to test whether the compositional ability of current models holds once a source image fixes the domain. We ask four questions:

- **Q1:** Does compositional ability drop from generation to editing (*generator-domain bias*)?
- **Q2:** Does the size of the drop depend on the source domain?
- **Q3:** Do models edit the source image, or redraw it in their preferred domain?
- **Q4:** When the edit is correct, do the added objects fit the scene?
- **Q5:** Do existing benchmarks rank models correctly?

Baselines. We evaluate eight open-source models. Three are editing-specific models: Qwen-Image-Edit Wu et al. (2025a), FLUX.1 Kontext Labs et al. (2025), and Step1X-Edit Liu et al. (2025). The others support both generation and editing: the unified models BAGEL Deng et al. (2025), SenseNova-U1 Diao et al. (2026), and OmniGen2 Wu et al. (2025b), and the generation-oriented FLUX.2 Klein Labs (2025). For SenseNova-U1, we evaluate two versions: SenseNova-SFT, and SenseNova-RL, which is further trained with reinforcement learning. Since both share the same base model, comparing them isolates the effect of RL post-training. For models that support generation, we also report their generation accuracy on the same prompts as a reference. Since Qwen-Image-Edit is built on Qwen-Image Wu et al. (2025a), we use Qwen-Image for its generation reference. Step1X-Edit and FLUX.1 Kontext do not support generation and therefore have no reference.

Generation and evaluation. For Texture and Landscape, we sample 60 prompts from GenEval (Ghosh et al., 2023). Editing prompts are written in the imperative form ("Add exactly ⟨count⟩ ⟨object⟩ to the image"), and generation prompts keep the original form ("a photo of ⟨object⟩") with the same explicit constraints on count (see Table 10 in Appendix). Because these constraints make the prompts stricter and we score outputs with Soft-TIFA instead of GenEval's de-

tector, our generation accuracies are not directly comparable to previously reported GenEval scores; several models score lower under the added "exactly" constraint. We use 3 random seeds.

Q1: Does compositional ability drop from generation to editing? We give each model the same prompt twice, once without a source image (generation) and once as an edit on an image, and compare their Soft-TIFA accuracy. Even though Texture sources contain no objects, every model with a generation reference falls below it on every texture (Table 2) and on every Landscape style (Table 3). Compositional ability therefore drops once the domain is fixed, as *generator-domain bias* predicts.

Table 2: **Texture subset.** Editing with "Add X" on textured images: Soft-TIFA (Acc; gm $\times 100$, ↑) and background MSE vs. R' (MSE; $\times 10^2$, ↓); 60 prompts $\times$ 5 seeds per cell. †Generation row uses Qwen-Image.

	BAGEL		Step-1X-Edit		FLUX.1 Kontext		Qwen-Image-Edit		OmniGen2		SenseNova-SFT		SenseNova-RL		FLUX2	
Texture	Acc	MSE	Acc	MSE	Acc	MSE	Acc	MSE	Acc	MSE	Acc	MSE	Acc	MSE	Acc	MSE
Generation	93.3	–	–	–	–	–	97.0†	–	88.78	–	98.0	–	98.5	–	95.24	–
white	77.1	2.0	74.3	58.0	69.0	0.3	58.1	189.3	63.8	61.2	94.7	0.7	94.5	113.7	53.17	0.6
black	75.9	339.6	79.7	141.0	64.4	50.8	68.0	223.6	61.9	310.3	90.9	35.3	94.4	235.6	52.25	0.7
gray	75.5	3.1	76.4	18.4	64.8	16.3	63.5	46.2	65.8	24.4	92.0	106.6	95.1	10.3	53.44	0.5
brick	62.5	1.0	67.1	12.2	65.6	8.1	85.6	4.8	68.3	23.1	82.8	2.2	90.8	2.6	70.3	1.9
wood	78.1	2.2	71.1	13.7	61.0	1.5	84.6	16.9	68.4	9.0	88.1	8.2	92.0	8.7	64.19	2.1
concrete	73.0	1.3	73.6	6.9	49.9	0.4	82.0	3.6	69.3	21.7	87.5	4.6	92.6	9.0	65.31	2.5
cracked earth	68.3	3.2	73.9	38.1	60.5	2.2	78.7	28.5	63.8	28.2	79.7	9.6	87.8	10.3	64.55	4.6
fabric	65.3	1.7	74.2	8.1	69.0	2.9	82.1	5.1	62.9	18.4	85.1	6.0	94.5	6.3	60.13	2.3
slate	71.0	2.0	75.1	3.3	47.2	0.8	82.5	7.9	59.7	68.8	86.0	2.0	90.4	3.6	56.02	1.0
marble	69.2	1.3	72.5	4.5	59.5	0.2	84.1	3.3	62.1	27.88	86.5	2.6	91.8	16.4	63.66	1.0
metal	55.9	1.6	70.8	23.5	61.8	9.6	80.4	35.7	55.9	18.5	83.0	3.9	87.1	4.5	62.55	1.4
paper	81.6	5.8	75.5	6.2	68.1	0.3	82.6	1.5	62.9	38.1	91.3	1.5	92.7	44.6	65.57	0.5

Q2: Does the size of the drop depend on the source domain? We compare accuracy across sources of increasing complexity (Texture and Landscape) and across styles, and on Scene we compare restyled sources with the original photos. For six of eight models, accuracy is lower on Landscape than on Texture: once the source contains a scene, composing within it becomes harder. BAGEL is affected most, dropping from 93.3 in generation to 71.1 on Texture and 37.3 on Landscape, while Qwen-Image-Edit barely changes (77.7 vs. 75.1). Within each subset, accuracy also depends on the texture (BAGEL: 55.9 on metal, 81.6 on paper) and on the style, where photorealistic sources are not the easiest and the hardest style differs between models (Tables 2 and 3). On Scene, every restyled source lowers semantic consistency for every model, most for cartoon and sketch (Table 4). The drop depends strongly on the source domain, and differs for each model.

Table 3: **Landscape.** Soft-TIFA (Acc; gm $\times 100$, ↑), background MSE (MSE; $\times 10^2$, ↓), and Context score (Ctx, ↑) on the Landscape subset, averaged per style. Each style averages its original images and the images transferred into it from the other four styles. †Generation row uses Qwen-Image, not Qwen-Image-Edit.

	BAGEL			Step-1X-Edit			FLUX.1 Kontext			Qwen-Image-Edit			OmniGen2			SenseNova-SFT			SenseNova-RL			FLUX2		
Style	Acc	MSE	Ctx	Acc	MSE	Ctx	Acc	MSE	Ctx	Acc	MSE	Ctx	Acc	MSE	Ctx	Acc	MSE	Ctx	Acc	MSE	Ctx	Acc	MSE	Ctx
Generation	93.3	–	–	–	–	–	–	–	–	97.0†	–	–	88.78	–	–	98.0	–	–	98.5	–	–	95.24	–	–
photo	41.9	0.4	4.74	58.5	4.9	5.95	46.1	4.1	5.95	75.2	2.8	5.78	69.5	5.6	4.61	82.7	1.4	5.52	87.8	2.2	5.54	60.7	1.0	5.92
cartoon	47.9	0.3	6.34	65.0	3.3	6.75	58.8	5.4	6.62	75.4	6.8	6.85	65.0	16.7	6.36	85.4	0.6	6.52	87.8	0.9	6.64	62.7	4.5	6.95
grayscale	29.4	0.3	5.63	65.3	1.9	5.79	42.5	0.7	5.31	74.9	1.6	5.87	62.4	27.1	5.24	82.0	1.0	5.48	87.5	1.9	5.54	61.0	0.5	5.88
painting	31.9	0.3	5.51	62.3	2.9	6.14	48.3	2.9	5.73	76.4	2.1	6.15	68.2	6.0	5.32	84.2	0.9	6.03	86.6	2.0	5.93	62.8	0.8	6.11
sketch	35.4	0.9	5.58	59.6	4.9	6.46	45.9	1.8	5.59	73.4	7.2	6.57	66.2	19.7	5.67	81.3	2.1	6.46	88.0	2.4	6.36	59.2	4.3	6.88

Q3: Do models edit the source image, or redraw it in their preferred domain? If a model edits the source, everything outside the added objects should stay unchanged. We therefore measure the MSE between the output and the source outside the added objects: a low value means the source was kept, a high value means it was redrawn. On black, background MSE is the highest of all textures for six of eight models, reaching 339.6 for BAGEL while its accuracy remains 75.9: rather than editing the source, the model regenerates the background in its preferred style and adds the objects there (Table 2). Step1X-Edit, Qwen-Image-Edit, and OmniGen2, for instance, turn the black canvas into a beach (Figure 14). The SenseNova pair shows that RL post-training can make this worse: SenseNova-RL regenerates the white background (MSE 113.7), while SenseNova-SFT preserves it (0.7) at similar accuracy (94.5 vs. 94.7). Models therefore often do not edit the source, but redraw it in the domain they prefer.

Q4: When the edit is correct, do the added objects fit the scene? To survey how well the added objects in GenEval-Edit are integrated within the context, we score Landscape edits with our context score, which rates on a 1–10 scale how naturally the added objects are integrated into the scene (support, scale, perspective, lighting). A high accuracy does not imply a high context score (Table 3): SenseNova-RL has the highest accuracy but only the fifth-highest context score (6.00), whereas FLUX.2, sixth in accuracy, has the highest context score (6.35). OmniGen2 has the lowest context score (5.44) and also changes the Landscape background the most (MSE up to 27.1). Placing the requested objects correctly is therefore not the same as integrating them into a scene the model did not choose. As a sanity check, a single annotator rated the integration of 77 edits per style on the same 1–10 scale; ranking the models by these ratings gives nearly the same order as the context score (Spearman $\rho = 0.90$, Figure 7), although with one annotator this is indicative only.

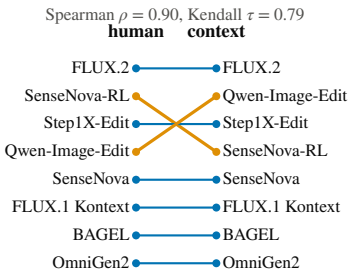


Figure 7: **Model ranks by a human annotator (left) and VLM judge (right).**

Q5: Do existing benchmarks rank models correctly? Existing benchmarks rank models either by generating from a prompt or by editing photographs. We compare these rankings with the rankings on DOMAIN-EDIT (Figure 8). Generation accuracy does not predict editing: BAGEL outperforms OmniGen2 in generation (93.3 vs. 88.8), yet OmniGen2 is far stronger on Landscape (66.3 vs. 37.3). Editing photographs does not predict editing other styles: SenseNova-RL achieves the highest semantic consistency on the original GEdit-Bench (8.02), but Qwen-Image-Edit is the strongest on Scene restyled sources (7.50). Accuracy also does not predict the context score: SenseNova-RL ranks first in Landscape accuracy but fifth in context score, and FLUX.2 sixth and first. Rankings from generation or photo-editing benchmarks therefore do not carry over once the model can no longer choose the domain.

Table 4: **Scene.** VIEScore averaged over the five core editing tasks, per input style. Colors are scaled separately for SC and PQ.

	BAGEL		Step-1X-Edit		FLUX.1 Kontext		Qwen-Image-Edit		OmniGen2		SenseNova-SFT		SenseNova-RL		FLUX2	
Style	SC	PQ	SC	PQ	SC	PQ	SC	PQ	SC	PQ	SC	PQ	SC	PQ	SC	PQ
photo (original)	7.53	7.12	7.87	7.58	6.40	7.47	7.89	7.60	7.52	7.14	7.99	7.50	8.02	7.52	7.89	7.60
cartoon	7.00	7.59	7.04	7.64	5.99	7.65	7.14	7.74	6.77	7.59	7.36	7.54	7.27	7.66	7.15	7.69
grayscale	6.96	7.12	7.43	7.43	6.30	7.33	7.64	7.53	6.99	7.19	7.64	7.35	7.61	7.35	7.67	7.40
painting	6.84	7.16	7.53	7.49	6.28	7.45	7.71	7.57	6.94	7.03	7.49	7.39	7.72	7.36	7.40	7.59
sketch	6.65	7.45	7.28	7.66	6.10	7.68	7.52	7.82	6.67	7.13	7.17	7.60	7.16	7.61	7.37	7.75

6 CONCLUSION

Compositional generation benchmarks are approaching saturation, yet generative models have been shown to favor narrow domains in layout and style. Because standard benchmarks let the model choose its own domain, they cannot tell whether high scores reflect general compositional ability or performance within that preferred domain. We introduced DOMAIN-EDIT, which fixes the domain through a source image and evaluates compositional editing on three subsets of increasing visual diversity: Texture, Landscape, and Scene. Across eight open-source models, we find that performance drops once the model can no longer choose its domain. Every model falls below its own generation accuracy, and generation scores fail to predict editing performance, with rankings changing across subsets and styles. Which styles are hard differs between models and subsets rather than following a single pattern, consistent with each model having its own preferred domain.

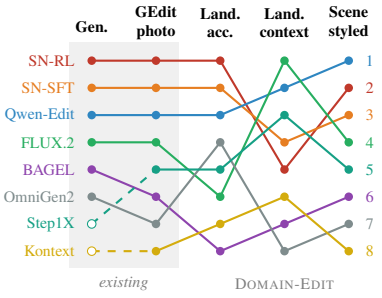


Figure 8: **Model rankings change across benchmarks.** Rank of each model (1 = best) under generation accuracy and GEdit-Bench on photos (shaded, existing benchmarks) and under DOMAIN-EDIT: Landscape accuracy, Landscape context score, and Scene on restyled sources. Step1X-Edit and FLUX.1 Kontext cannot generate (hollow markers).

AI Disclosure

We used generative AI tools in three ways. First, the stylized source images of our benchmark were generated with Nano Banana Pro and Nano Banana 2 (Section A). Second, VLMs (Qwen3.5-9B, Qwen3-VL-8B, Qwen2.5-VL-72B) serve as detectors and judges in our evaluation pipeline, as described in Section A. Third, we used a coding assistants to polish the writing. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

References

Krzysztof Adamkiewicz, Brian Bernhard Moser, Stanislav Frolov, Tobias Christian Nauen, Federico Raue, and Andreas Dengel. When pretty isn't useful: Investigating why modern text-to-image models fail as reliable training data generators, 2026. URL https://arxiv.org/abs/2602.19946.

Samyadeep Basu, Mehrdad Saberi, Shweta Bhardwaj, Atoosa Malemir Chegini, Daniela Massiceti, Maziar Sanjabi, Shell Xu Hu, and Soheil Feizi. Editval: Benchmarking diffusion based text-guided image editing methods. *arXiv preprint arXiv:2310.02426*, 2023.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18392–18402. IEEE, 2023.

Chubin Chen, Sujie Hu, Jiashu Zhu, Meiqi Wu, Jintao Chen, Yanxun Li, Nisha Huang, Chengyu Fang, Jiahong Wu, Xiangxiang Chu, et al. Taming preference mode collapse via directional decoupling alignment in diffusion reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12775–12786, 2026.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining, 2025. URL https://arxiv.org/abs/2505.14683.

Haiwen Diao, Penghao Wu, Hanming Deng, Jiahao Wang, Shihao Bai, Silei Wu, Weichen Fan, Wenjie Ye, Wenwen Tong, Xiangyu Fan, et al. Sensenova-u1: Unifying multimodal understanding and generation with neo-unify architecture. *arXiv preprint arXiv:2605.12500*, 2026.

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.

Saba Fathi, Fardin Ayar, Maryam Abdolali, Ehsan Javanmardi, Manabu Tsukada, and Mahdi Javanmardi. How do diffusion classifiers decide? a bias-centric evaluation. *arXiv preprint arXiv:2607.03831*, 2026.

Junyao Gao, Sibo Liu, Jiaxing Li, Yanan Sun, Yuanpeng Tu, Fei Shen, Weidong Zhang, Cairong Zhao, and Jun Zhang. Megastyle: Constructing diverse and scalable style dataset via consistent text-to-image style mapping. *arXiv preprint arXiv:2604.08364*, 2026.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36: 52132–52152, 2023.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023.

Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench++: An enhanced and comprehensive benchmark for compositional text-to-image generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):3563–3579, 2025.

Yujin Jeong, Arnas Uselis, Seong Joon Oh, and Anna Rohrbach. Diffusion classifiers understand compositionality, but conditions apply. *arXiv preprint arXiv:2505.17955*, 2:9, 2025.

Qing Jiang, Junan Huo, Xingyu Chen, Yuda Xiong, Zhaoyang Zeng, Yihao Chen, Tianhe Ren, Junzhi Yu, and Lei Zhang. Detect anything via next point prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25472–25483, 2026.

Amita Kamath, Kai-Wei Chang, Ranjay Krishna, Luke Zettlemoyer, Yushi Hu, and Marjan Ghazvininejad. Geneval 2: Addressing benchmark drift in text-to-image evaluation. *arXiv preprint arXiv:2512.16853*, 2025.

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, 2024.

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, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Wen Li, Muyuan Fang, Cheng Zou, Biao Gong, Ruobing Zheng, Meng Wang, Jingdong Chen, and Ming Yang. Styletokenizer: Defining image style by a single instance for controlling diffusion models. In *European conference on computer vision*, pp. 110–126. 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.

Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7176–7185. PMLR, 2020. URL `https://proceedings.mlr.press/v119/naeem20a.html`.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, volume 2024, pp. 1862–1874, 2024.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL `https://qwen.ai/blog?id=qwen3.5`.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. Emu edit: Precise image editing via recognition and generation tasks. In *2024 ieee/cvf conference on computer vision and pattern recognition (cvpr)*, pp. 8871–8879. IEEE, 2024.

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

Arnas Uselis, Andrea Dittadi, and Seong Joon Oh. Compositional generalization requires linear, orthogonal representations in vision embedding models. *arXiv preprint arXiv:2602.24264*, 2026.

Ye Wang, Ruiqi Liu, Jiang Lin, Fei Liu, Zili Yi, Yilin Wang, and Rui Ma. Omnistyle: Filtering high quality style transfer data at scale. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7847–7856. IEEE, 2025.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025a. URL https://arxiv.org/abs/2508.02324.

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, Ze Liu, Ziyi Xia, Chaofan Li, Haoge Deng, Jiahao Wang, Kun Luo, Bo Zhang, Defu Lian, Xinlong Wang, Zhongyuan Wang, Tiejun Huang, and Zheng Liu. Omnigen2: Towards instruction-aligned multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025b.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. In *International Conference on Learning Representations*, volume 2025, pp. 28240–28264, 2025.

Yihua Zhang, Chongyu Fan, Yimeng Zhang, Yuguang Yao, Jinghan Jia, Jiancheng Liu, Gaoyuan Zhang, Gaowen Liu, Ramana Rao Kompella, Xiaoming Liu, et al. Unlearncanvas: Stylized image dataset for enhanced machine unlearning evaluation in diffusion models. *arXiv preprint arXiv:2402.11846*, 2024.

Shihao Zhu, Ziheng Ouyang, Yijia Kang, Qilong Wang, Mi Zhou, Bo Li, Ming-Ming Cheng, and Qibin Hou. Mixture of style experts for diverse image stylization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30500–30510, 2026.

APPENDIX

Table of Contents

A Benchmark details **14**

A.1 Preliminaries 14

A.2 Dataset details 14

A.3 Generation details 15

A.4 Evalution details 17

B Full results **18**

B.1 Landscape 18

B.2 Scene 19

C Additional experiments **20**

C.1 Ablation study of prompts 20

C.2 Where objects land: generation vs. editing 20

C.3 Texture qualitative examples 22

C.4 Landscape qualitative examples 34

C.5 Scene qualitative examples 59

AI DISCLOSURE

We used generative AI tools in three ways. First, the stylized source images of our benchmark were generated with Nano Banana Pro and Nano Banana 2 (Section A). Second, VLMs (Qwen3.5-9B, Qwen3-VL-8B, Qwen2.5-VL-72B) serve as detectors and judges in our evaluation pipeline, as described in Section A. Third, we used a coding assistants to polish the writing. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

A BENCHMARK DETAILS

A.1 PRELIMINARIES

Our benchmark is built upon two existing benchmarks. We detail them below.

GenEval: generation benchmark. GenEval Ghosh et al. (2023) evaluates compositional generation, where images are sampled as $x \sim p(x \mid y)$ from templated prompts over six tasks: single object, two objects, counting, colors, position, and color attribution. Each generated image is verified with an object detector, combined with a CLIP model for color classification. GenEval 2 Kamath et al. (2025) extends it with more challenging prompts and introduces Soft-TIFA, which decomposes each prompt into atomic questions (e.g., whether an object is present, or whether it has the right color) and answers them with a vision-language model (VLM), making evaluation more robust than detector-based verification.

GEdit-Bench: editing benchmark. GEdit-Bench Liu et al. (2025) evaluates instruction-based editing on 606 real-world source images paired with editing instructions across 11 task categories. Outputs are scored with VIEScore Ku et al. (2024), where a VLM judge rates semantic consistency (SC), i.e., whether the edit follows the instruction, and perceptual quality (PQ), i.e., whether the result looks natural.

A.2 DATASET DETAILS

Table 5 summarizes the source content, construction, size, instructions, and evaluation of each DOMAIN-EDIT subset.

Table 5: **Overview of the three DOMAIN-EDIT subsets.** All subsets keep the instruction fixed and vary only the source image; they differ in what the source contains, with visual diversity increasing from left to right. Scene images and instructions are taken from GEdit-Bench Liu et al. (2025).

Subset	Texture	Landscape	Scene
Content	plain textures	background-only scenes	object-centric photos
Origin	collected	collected + style transfer	GEdit-Bench + style transfer
Variations	12 textures	5 styles	5 styles
# sources	12	$5 \times 5 = 25$	$205 \times 5 = 1{,}025$
Instructions	GenEval-based	GenEval-based	GEdit-Bench
Judge	Qwen3-VL-8B	Qwen3-VL-8B	Qwen2.5-VL-72B
Alignment	Soft-TIFA	Soft-TIFA	VIEScore SC
Preservation	BG MSE	BG MSE	VIEScore PQ
Context	–	Context score	–

Texture subset details. Sources are nine textures from the web and three monotone (black, gray, and white) backgrounds, containing no objects, horizon, or ground plane. They serve as a layout-free baseline (more examples in Figure 9). Instructions are derived from GenEval: 60 prompts covering its six tasks (single object, two objects, counting, colors, position, and color attribution), with 10 prompts per task. Prompts are rewritten from "a photo of [object]" into the imperative "add [object] to the image", since with the original phrasing models often ignored the source and regenerated the background, even on simple backgrounds (see Table 10 for a comparison of prompt formats). Each prompt also states the exact number of objects, forbids duplicates, and explicitly specifies spatial



Figure 9: **Source images of the Texture subset.** Nine textures collected from Unsplash (concrete, marble, slate, cracked earth, fabric, metal, wood, brick, and paper) and three monotone backgrounds (black, white, and gray). None of the sources contains objects, a horizon, or a ground plane, so they serve as a layout-free baseline.

relations. This removes ambiguity for the judge: for instance, it is unclear whether an image with two dogs satisfies "add a dog". Target objects are balanced between those that fit the context (e.g., umbrella, sports ball, bottle) and those that do not (e.g., notebook, sink, oven).

Landscape subset details. We collect background-only scenes, such as landscapes, from the web, each originally depicted in one of five styles: photorealistic, cartoon, painting, grayscale, and sketch. For each scene, we render the four styles it was not originally depicted in with Nano Banana Pro Team et al. (2023), yielding a complete grid in which every scene appears in every style. We use Nano Banana Pro because it is closed-source and not among the evaluated models, so no evaluated model is tested on its own outputs. Each style contains the same number of original and generated images, so differences across styles cannot be explained by some styles containing more generated images than others. More examples are in Figure 10.

Scene subset. We take the source images of GEdit-Bench, which are predominantly photographic and object-centric. We first classify the style of each source image with Qwen3.5-9B and find that about 90% are photorealistic (Figure 11). To avoid restyling images that are already stylized (e.g., turning a cartoon into another cartoon), we keep only the photorealistic images and render each in cartoon, painting, sketch, and grayscale with Nano Banana 2 Team et al. (2023), preserving content and layout. The original editing instructions are kept unchanged and applied to both the original and the restyled sources, so that only the source style varies.

A.3 Generation details

We use separate prompts for the generation baseline and for editing (Table 10). Generation prompts keep GenEval's form ("a photo of ⟨object⟩"), while editing prompts are rewritten in the imperative ("Add ⟨object⟩ to the image"). Both versions include the same explicit constraints on the number of objects, duplicates, and spatial relations, so that generation and editing are scored against identical criteria. These constraints also remove ambiguity for the judge: without them, it is unclear whether an image with two dogs satisfies "a dog".

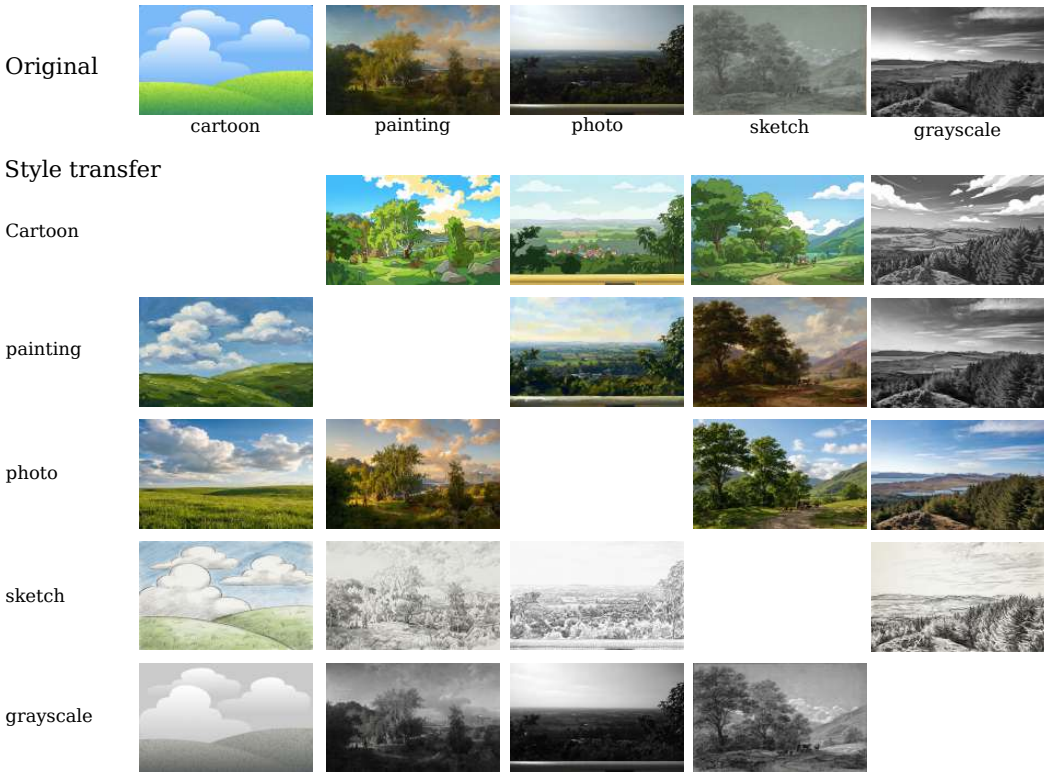


Figure 10: **Source images of the Landscape subset.** Each column is one scene. The top row shows the five collected originals, each in a different style. The rows below show each scene rendered in the remaining styles with Nano Banana Pro; the empty cell in each column corresponds to the scene's original style. Together, every scene appears in all five styles, giving $5 \times 5 = 25$ source images.

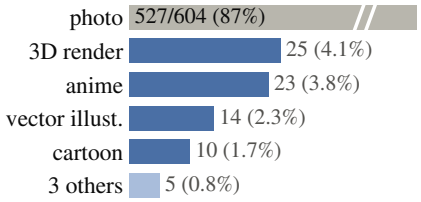


Figure 11: **Style distribution of GEdit-Bench source images.**

Task	Generation prompt	Editing prompt
`object`	a photo of one ⟨object⟩. Do not add any duplicate objects.	Add exactly one ⟨object⟩ to the image. Do not add any duplicate objects.
`count`	a photo of exactly ⟨count⟩ ⟨object⟩.	Add exactly ⟨count⟩ ⟨object⟩ to the image.
`two objects`	a photo of one ⟨object1⟩ and one ⟨object2⟩. Do not add any duplicate objects.	Add exactly two new objects to the image: one ⟨object1⟩ and one ⟨object2⟩. Do not add any duplicate objects.
`color`	a photo of one ⟨color⟩ ⟨object⟩. Do not add any duplicate objects.	Add exactly one ⟨color⟩ ⟨object⟩ to the image. Do not add any duplicate objects.
`position`	a photo of one ⟨object1⟩ and one ⟨object2⟩, with the ⟨object1⟩ directly to the left of the ⟨object2⟩, aligned horizontally. Do not add any duplicate objects.	Add exactly two new objects to the image: one ⟨object1⟩ and one ⟨object2⟩. The ⟨object1⟩ should be directly to the left of the ⟨object2⟩, aligned horizontally. Do not add any duplicate objects.
`color attribution`	a photo of one ⟨color1⟩ ⟨object1⟩ and one ⟨color2⟩ ⟨object2⟩. Do not add any duplicate objects.	Add exactly two new objects to the image: one ⟨color1⟩ ⟨object1⟩ and one ⟨color2⟩ ⟨object2⟩. Do not add any duplicate objects.

A.4 EVALUTION DETAILS

Semantic alignment. We use Soft-TIFA Kamath et al. (2025) as the judge, which decomposes each prompt into atomic queries on object presence, color, position, and count (Table 7) and answers each with a VLM (Qwen3-VL-8B) as a soft probability. Following Soft-TIFA, we take their geometric mean, so that a single violated condition, such as a duplicated or misplaced object, strongly lowers the score. To validate the judge, human annotators verified the atomic criteria on 6,300 edited images across styles and textures. For Scene, we follow GEdit-Bench Liu et al. (2025) and report the semantic consistency (SC) score of VIEScore, computed with Qwen2.5-VL-72B (AWQ).

Primitive	**Question Template**	**Target Answer**
`count`	*How many <object> are in the image?*	number word
`attribute`	*Are the <object> <color>?*	Yes
`object`	*Are there any <object> in the image?*	Yes
`position`	*Are the <object> <phrase> the <ref object>?*	Yes

Table 7: **Atomic Verification Templates in Soft-TIFA.** Decomposed primitive question templates, expected target answers, and condition criteria for inclusion in the atomic count metric.

Source preservation. A high semantic alignment score does not by itself mean the edit succeeded: a model can add the correct objects while regenerating the background in its preferred style, which defeats the purpose of fixing the domain. For Texture and Landscape, we therefore measure how much the background changes. We localize the target objects with Rex-Omni Jiang et al. (2026), queried only with the object classes named in the prompt, remove near-duplicate boxes (largest first, IoU ≥ 0.9), and use the union of the remaining boxes as the object mask. We then compute the MSE (0–255, averaged over RGB) outside this mask between the output and the source passed through the same resizing and VAE reconstruction as the evaluated editor, which removes the error introduced by encoding alone; this reference is model-specific, since different editors preprocess the source differently. Samples with no detected target are excluded, and their rate is reported separately; lower values indicate better background preservation. For Scene, we follow GEdit-Bench and report the perceptual quality (PQ) score of VIEScore, which rates how natural and artifact-free the output looks. While background MSE measures fidelity to the source and PQ measures the quality of the output itself, both capture whether the edit is carried out without degrading the image.

Context score. Prompt compliance does not imply visual plausibility: a requested bird may be present, but float in mid-air, and an inserted object may have inconsistent scale, lighting, contact, perspective, or boundaries. For Landscape, we measure this in three stages. First, Qwen3.5-9B receives the edited image, the prompt, and the requested objects, and produces a compact scene graph (at most eight prominent objects and their relations) together with an assessment of each requested object. It explicitly checks for floating objects, implausible support or contact, pasted-on boundaries and halos, mismatched lighting and shadows, scale and perspective, occlusion, texture and style consistency, and spatial placement, and is instructed not to penalize an image merely for being stylized. Second, the objects are localized with Rex-Omni Jiang et al. (2026), which is the only source of bounding boxes; boxes predicted by Qwen are discarded. Third, a second Qwen3.5-9B query receives the image and this analysis and rates integration on a 1–10 scale, emitting a single constrained token. From the token probabilities we compute the expected score $\sum_{k=1}^{10} k\,P(\text{score} = k)$, which captures the judge's uncertainty and is the value we report. Context scores are reported either over all images or over the prompt-compliant subset, in which every Soft-TIFA atom has probability at least 0.8; we state which population is used in each case.

B FULL RESULTS

This section reports the full results behind the per-style averages in the main text.

B.1 LANDSCAPE

Table 8 lists Soft-TIFA accuracy and background MSE for all 25 source images of the Landscape subset. Each row corresponds to one combination of original and target style: a single label (e.g., *cartoon*) denotes an original image, and *source* $\rightarrow$ *target* denotes an image transferred from the source style into the target style. The per-style averages in Table 3 are computed by averaging, for each target style, its original image and the four images transferred into it.

Table 8: Soft-TIFA (Acc; gm $\times 100$, $\uparrow$) and background MSE (MSE; $\times 10^2$, $\downarrow$) by style and model on stylized sources. Row labels read *source style* $\rightarrow$ *transferred style*; a single label is an untransferred source. †Generation row uses Qwen-Image, not Qwen-Image-Edit.

	BAGEL		**Step-1X-Edit**		**FLUX.1 Kontext**		**Qwen-Image-Edit**		**OmniGen2**		**SenseNova-SFT**		**SenseNova-RL**		**FLUX2**	
Style	Acc	MSE	Acc	MSE	Acc	MSE	Acc	MSE	Acc	MSE	Acc	MSE	Acc	MSE	Acc	MSE
Generation	93.3	–	–	–	–	–	97.0†	–	88.78	–	98.0	–	98.5	–	95.24	–
cartoon	76.9	0.2	74.5	0.2	71.3	7.6	85.9	2.7	64.71	56.2	88.95	0.6	91.05	1.0	68.01	0.5
cartoon $\rightarrow$ grayscale	71.2	0.2	64.0	0.6	67.6	1.3	81.5	0.9	58.68	36.4	84.58	0.2	89.09	0.5	69.54	0.4
cartoon $\rightarrow$ painting	61.9	0.5	52.1	3.2	70.8	10.6	84.0	3.0	70.39	8.2	87.66	0.6	89.5	0.9	55.04	1.1
cartoon $\rightarrow$ photo	74.1	0.4	62.9	3.0	66.3	14.2	81.1	4.6	70.39	7.6	85.14	1.1	91.92	1.3	69.1	1.0
cartoon $\rightarrow$ sketch	70.0	0.5	46.8	1.8	62.5	6.4	77.1	7.6	70.96	22.9	86.28	0.9	89.92	1.3	65.17	0.9
grayscale	28.8	0.3	63.8	2.0	33.7	0.1	72.9	1.0	56.73	58.9	83.09	1.0	89.07	1.5	61.65	0.3
grayscale $\rightarrow$ cartoon	32.6	0.3	63.6	0.7	40.5	0.3	65.8	1.2	47.95	11.8	79.81	0.5	84.01	1.2	57.16	3.8
grayscale $\rightarrow$ painting	23.6	0.4	53.3	3.0	35.1	0.2	78.1	3.4	51.65	8.7	80.65	1.4	84.46	1.5	64.28	1.9
grayscale $\rightarrow$ photo	46.4	0.4	56.0	2.4	29.9	0.5	76.4	2.9	64.02	2.2	83.98	0.9	86.98	1.7	62.36	0.9
grayscale $\rightarrow$ sketch	33.9	1.0	67.4	3.7	43.3	0.7	65.2	10.6	62.11	16.2	79.63	3.3	88.62	3.9	59.07	9.9
painting	12.4	0.3	65.8	0.6	34.7	0.3	71.0	2.5	69.13	4.6	78.37	1.0	83.65	1.1	65.74	0.3
painting $\rightarrow$ cartoon	17.7	0.3	51.7	2.3	66.5	9.2	73.7	6.5	62.67	3.4	86.75	0.5	89.8	0.8	68.48	6.1
painting $\rightarrow$ grayscale	22.5	0.2	69.3	0.5	42.8	0.3	66.8	1.4	66.96	19.4	77.11	0.8	85.82	1.0	64.26	1.1
painting $\rightarrow$ photo	23.5	0.3	51.4	2.9	43.1	3.1	72.1	2.0	70.73	3.3	81.91	1.7	85.17	2.3	60.85	2.0
painting $\rightarrow$ sketch	27.0	1.1	51.4	3.6	41.6	1.0	67.9	9.1	61.51	23.5	83.6	2.6	89.02	2.5	56.34	8.0
photo	21.7	0.5	66.7	3.3	37.5	2.2	75.4	3.3	71.71	8.7	80.12	1.1	88.16	3.5	44.71	0.4
photo $\rightarrow$ cartoon	61.7	0.3	67.6	6.8	55.3	8.4	82.4	20.0	73.59	10.2	87.05	0.6	87.25	0.8	61.05	6.9
photo $\rightarrow$ grayscale	15.6	0.3	63.1	2.9	35.4	1.4	81.3	3.6	67.05	13.8	85.68	1.1	87.08	4.5	53.35	0.5
photo $\rightarrow$ painting	55.0	0.2	76.2	4.4	48.4	3.1	75.0	1.2	75.7	6.4	92.69	1.3	87.91	5.8	64.04	0.6
photo $\rightarrow$ sketch	34.7	1.6	70.6	13.3	38.9	0.7	81.8	6.6	72.9	32.2	83.0	2.7	86.91	3.2	59.08	2.3
sketch	11.6	0.2	61.9	2.2	43.2	0.3	75.2	2.2	63.64	3.5	73.92	1.0	85.62	1.1	56.51	0.2
sketch $\rightarrow$ cartoon	50.6	0.2	67.8	6.7	60.2	1.5	69.4	3.8	76.32	2.0	84.43	0.6	87.0	0.7	58.71	5.0
sketch $\rightarrow$ grayscale	9.0	0.5	66.3	3.4	33.0	0.5	72.2	1.0	62.71	7.2	79.4	1.8	86.3	1.8	56.01	0.3
sketch $\rightarrow$ painting	6.8	0.1	64.1	3.1	52.3	0.3	74.1	0.4	74.34	2.2	81.65	0.4	87.61	0.6	64.69	0.3
sketch $\rightarrow$ photo	44.0	0.6	55.7	13.0	53.9	0.6	71.2	1.0	70.57	6.0	82.1	2.3	86.89	2.2	66.68	0.7

B.2 SCENE

Table 9 reports VIEScore for each of the five core editing tasks of GEdit-Bench (subject addition, subject removal, subject replacement, background change, and motion change) and each source style. The averages over these five tasks correspond to Table 4 in the main text.

Table 9: GEdit-Bench VIEScore (SC / PQ) per editing task and input style.

Metric	Model	Photo (Orig.)	Cart.	Gray	Paint	Sketch	Mean
subject-add (SC↑ / PQ↑)	BAGEL	7.56 / 7.44	7.60 / 7.70	7.27 / 7.31	7.38 / 7.36	6.72 / 7.64	7.31 / 7.49
	Step-1X	7.77 / 7.67	7.30 / 7.74	7.24 / 7.44	7.38 / 7.49	7.68 / 7.70	7.48 / 7.61
	FLUX.1 Kontext	7.04 / 7.60	6.57 / 7.53	7.36 / 7.42	6.83 / 7.51	6.70 / 7.64	6.90 / 7.54
	Qwen-Image-Edit	8.08 / 7.58	7.47 / 7.77	7.93 / 7.53	7.72 / 7.53	7.79 /7.79	7.80 / 7.64
	Omnigen2	7.65 / 7.33	6.98 / 7.70	7.27 / 7.33	7.15 / 7.13	6.47 / 7.62	7.10 / 7.42
	SenseNova-SFT	7.88 / 7.60	7.72 / 7.70	7.69 / 7.44	7.83 / 7.36	7.47 / 7.68	7.72 / 7.56
	SenseNova-RL	8.08 / 7.65	7.85 / 7.64	7.96 / 7.29	7.83 / 7.36	7.66 / 7.64	7.88 / 7.51
	FLUX.2.Klein	8.10 / 7.62	7.85 / 7.72	7.98 / 7.56	7.79 / 7.45	7.72 / 7.74	7.89 / 7.62
subject-remove (SC↑ / PQ↑)	BAGEL	8.19 / 7.11	7.02 / 7.72	7.19 / 7.00	7.02 / 7.15	6.77 / 7.30	7.24 / 7.26
	Step-1X	8.47 / 7.64	6.85 / 7.66	7.62 / 7.23	7.79 / 7.53	7.00 / 7.51	7.54 / 7.51
	FLUX.1 Kontext	6.98 / 7.49	6.23 / 7.68	6.13 / 7.34	6.06 / 7.28	6.45 / 7.68	6.37 / 7.49
	Qwen-Image-Edit	8.13 / 7.55	6.68 / 7.72	7.28 / 7.36	7.45 / 7.45	7.34 /7.81	7.37 / 7.58
	Omnigen2	7.28 / 7.11	6.30 / 7.66	6.30 / 6.81	5.70 / 6.83	6.36 / 6.68	6.39 / 7.02
	SenseNova-SFT	8.23 / 7.47	7.17 / 7.45	7.55 / 7.15	6.70 / 7.32	6.74 / 7.30	7.28 / 7.34
	SenseNova-RL	8.02 / 7.51	6.89 / 7.72	7.15 / 7.19	7.51 / 7.26	6.06 / 7.62	7.13 / 7.46
	FLUX.2.Klein	7.83 / 7.60	6.57 / 7.85	7.21 / 7.36	6.43 / 7.45	7.09 / 7.64	7.03 / 7.58
subject-replace (SC↑ / PQ↑)	BAGEL	7.35 / 7.06	6.98 / 7.48	6.96 / 7.15	7.02 / 7.06	6.56 / 7.54	6.97 / 7.26
	Step-1X	7.48 / 7.54	6.81 / 7.52	7.57 / 7.48	7.43 / 7.38	7.31 / 7.81	7.32 / 7.55
	FLUX.1 Kontext	6.62 / 7.44	5.77 / 7.54	6.15 / 7.39	6.49 / 7.47	5.96 / 7.73	6.20 / 7.51
	Qwen-Image-Edit	8.00 / 7.54	7.31 / 7.62	7.98 / 7.46	7.83 / 7.57	7.56 /7.75	7.74 / 7.59
	Omnigen2	7.71 / 7.08	6.79 / 7.48	7.57 / 7.24	7.47 / 6.98	7.04 / 7.29	7.32 / 7.21
	SenseNova-SFT	8.08 / 7.38	7.17 / 7.52	8.00 / 7.41	7.79 / 7.32	7.46 / 7.73	7.70 / 7.47
	SenseNova-RL	8.10 / 7.33	7.27 / 7.54	7.89 / 7.39	7.83 / 7.36	7.52 / 7.58	7.72 / 7.44
	FLUX.2.Klein	7.94 / 7.44	7.62 / 7.52	7.85 / 7.33	7.70 / 7.53	7.79 / 7.77	7.78 / 7.52
background-change (SC↑ / PQ↑)	BAGEL	7.50 / 7.15	7.06 / 7.39	7.18 / 7.06	7.24 / 7.24	6.88 / 7.18	7.17 / 7.20
	Step-1X	7.97 / 7.65	7.58 / 7.58	7.59 / 7.38	7.88 / 7.42	7.61 / 7.55	7.72 / 7.52
	FLUX.1 Kontext	6.71 / 7.41	6.15 / 7.79	7.15 / 7.09	6.94 / 7.42	6.18 / 7.67	6.63 / 7.48
	Qwen-Image-Edit	8.03 / 7.79	7.61 / 7.79	8.00 / 7.71	8.03 / 7.73	7.76 /7.85	7.88 / 7.77
	Omnigen2	7.82 / 6.88	7.30 / 7.52	7.09 / 7.21	7.82 / 7.15	7.00 / 6.91	7.41 / 7.13
	SenseNova-SFT	7.94 / 7.56	7.55 / 7.48	7.47 / 7.35	7.91 / 7.48	7.30 / 7.67	7.63 / 7.51
	SenseNova-RL	7.94 / 7.68	7.36 / 7.82	7.47 / 7.38	7.94 / 7.36	7.58 / 7.58	7.66 / 7.56
	FLUX.2.Klein	8.00 / 7.76	7.21 / 7.70	7.91 / 7.26	7.94 / 7.79	7.18 / 7.73	7.65 / 7.65
motion-change (SC↑ / PQ↑)	BAGEL	7.03 / 6.83	6.33 / 7.67	6.23 / 7.06	5.56 / 7.00	6.33 / 7.58	6.30 / 7.23
	Step-1X	7.67 / 7.39	6.64 / 7.72	7.11 / 7.60	7.19 / 7.61	6.81 / 7.72	7.08 / 7.61
	FLUX.1 Kontext	4.67 / 7.42	5.19 / 7.69	4.71 / 7.40	5.06 / 7.56	5.19 / 7.69	4.97 / 7.55
	Qwen-Image-Edit	7.19 / 7.53	6.64 / 7.81	7.00 / 7.57	7.50 / 7.56	7.17 /7.92	7.10 / 7.68
	Omnigen2	7.14 / 7.31	6.50 / 7.58	6.74 / 7.37	6.58 / 7.06	6.47 / 7.14	6.69 / 7.29
	SenseNova-SFT	7.83 / 7.47	7.19 / 7.53	7.49 / 7.40	7.22 / 7.47	6.89 / 7.64	7.32 / 7.50
	SenseNova-RL	7.94 / 7.42	6.97 / 7.56	7.57 / 7.49	7.47 / 7.47	6.97 / 7.61	7.39 / 7.51
	FLUX.2.Klein	7.56 / 7.56	6.50 / 7.67	7.40 / 7.51	7.17 / 7.75	7.06 / 7.86	7.14 / 7.67
Average (core-5) (SC↑ / PQ↑)	BAGEL	7.53 / 7.12	7.00 / 7.59	6.96 / 7.12	6.84 / 7.16	6.65 / 7.45	7.00 / 7.29
	Step-1X	7.87 / 7.58	7.04 / 7.64	7.43 / 7.43	7.53 / 7.49	7.28 / 7.66	7.43 / 7.56
	FLUX.1 Kontext	6.40 / 7.47	5.99 / 7.65	6.30 / 7.33	6.28 / 7.45	6.10 / 7.68	6.21 / 7.52
	Qwen-Image-Edit	7.89 / 7.60	7.14 / 7.74	7.64 / 7.53	7.71 / 7.57	7.52 /7.82	7.58 / 7.65
	Omnigen2	7.52 / 7.14	6.77 / 7.59	6.99 / 7.19	6.94 / 7.03	6.67 / 7.13	6.98 / 7.22
	SenseNova-SFT	7.99 / 7.50	7.36 / 7.54	7.64 / 7.35	7.49 / 7.39	7.17 / 7.60	7.53 / 7.48
	SenseNova-RL	8.02 / 7.52	7.27 / 7.66	7.61 / 7.35	7.72 / 7.36	7.16 / 7.61	7.55 / 7.50
	FLUX.2.Klein	7.89 / 7.60	7.15 / 7.69	7.67 / 7.40	7.40 / 7.59	7.37 / 7.75	7.50 / 7.61

C ADDITIONAL EXPERIMENTS

C.1 ABLATION STUDY OF PROMPTS

We compare editing with our imperative prompts ("Add X") against editing with GenEval's original generation-style prompts ("A photo of X"), using the same sources and judge. Generation-style prompts raise accuracy for every model, but they also raise background MSE for every model (**??**): models treat the prompt as a request to generate a new image and redraw the source instead of editing it. This is why we use imperative prompts throughout. Therefore, we kept "Add" as editing prompts.

Table 10: Effect of prompt phrasing, averaged over 12 textured sources (60 prompts $\times$ 5 seeds each). "Gen prompt" = "A photo of X", "Edit prompt" = "Add X". Generation-style prompts raise contents accuracy for every model but also raise background MSE: the gain comes from regenerating the source rather than editing it.

	Soft-TIFA $\uparrow$		**bg_mse** $(\times 10^2)\downarrow$	
Model	Edit prompt	Gen prompt	Edit prompt	Gen prompt
BAGEL	71.1	72.0	30.4	60.0
Step-1X-Edit	73.7	80.6	27.8	34.5
FLUX.1 Kontext	61.7	74.4	8.0	52.6
Qwen-Image-Edit	77.7	84.4	47.3	61.4
Omnigen2	63.7	87.6	54.2	87.8
SenseNova-SFT	87.3	95.4	15.3	85.5
SenseNova-RL	92.0	95.9	38.8	68.6
FLUX.2 Klein	60.9	94.7	1.6	43.4

C.2 WHERE OBJECTS LAND: GENERATION VS. EDITING

Figures 12 and 13 show where the added objects land when the same GenEval prompts are generated from scratch (text-to-image) and when they are added as edits to each of the five scenes. In generation, box centers concentrate in the center of the canvas, reflecting each model's preferred layout. In editing, however, they spread out or cluster in regions that depend on the scene, indicating that the source image, rather than the model's prior, determines where objects are placed.

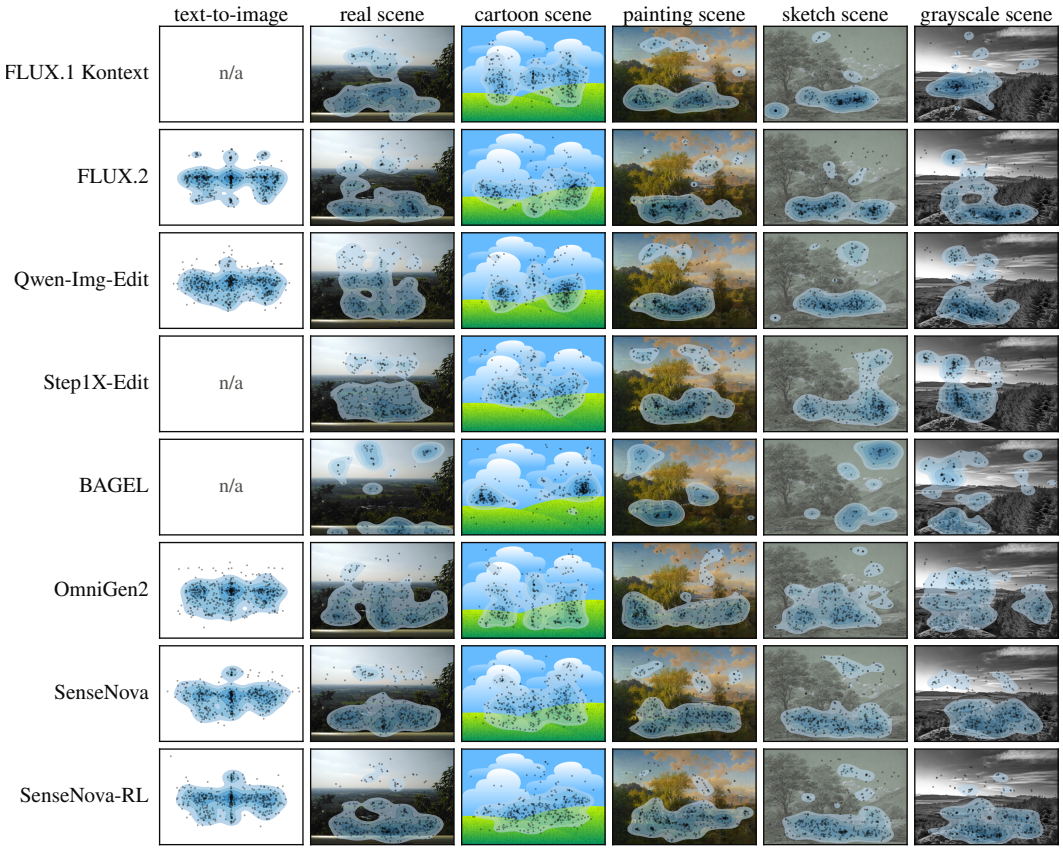


Figure 12: Box centres (dots) and their density (shading) per model, in text-to-image generation (blank canvas; n/a: no text-to-image run) and as edits on each scene (background shown). All tasks pooled.

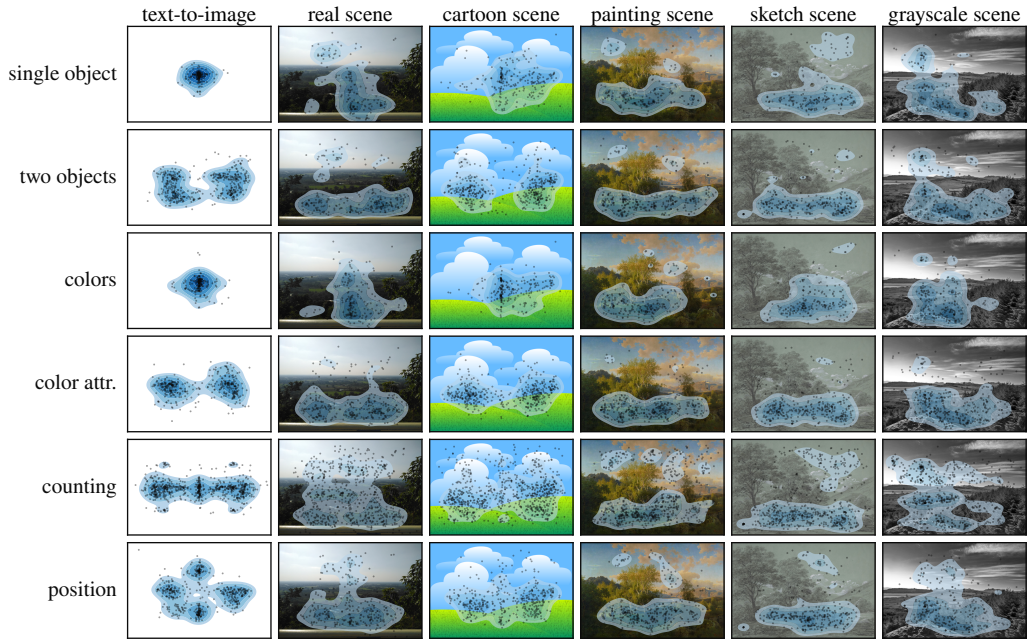


Figure 13: As Figure 12, per GenEval category, pooled over models.

C.3 TEXTURE QUALITATIVE EXAMPLES

Figures 14 to 25 show example edits on each of the twelve texture sources for all models. Each figure shows the source image and the outputs of each model for the same prompts.

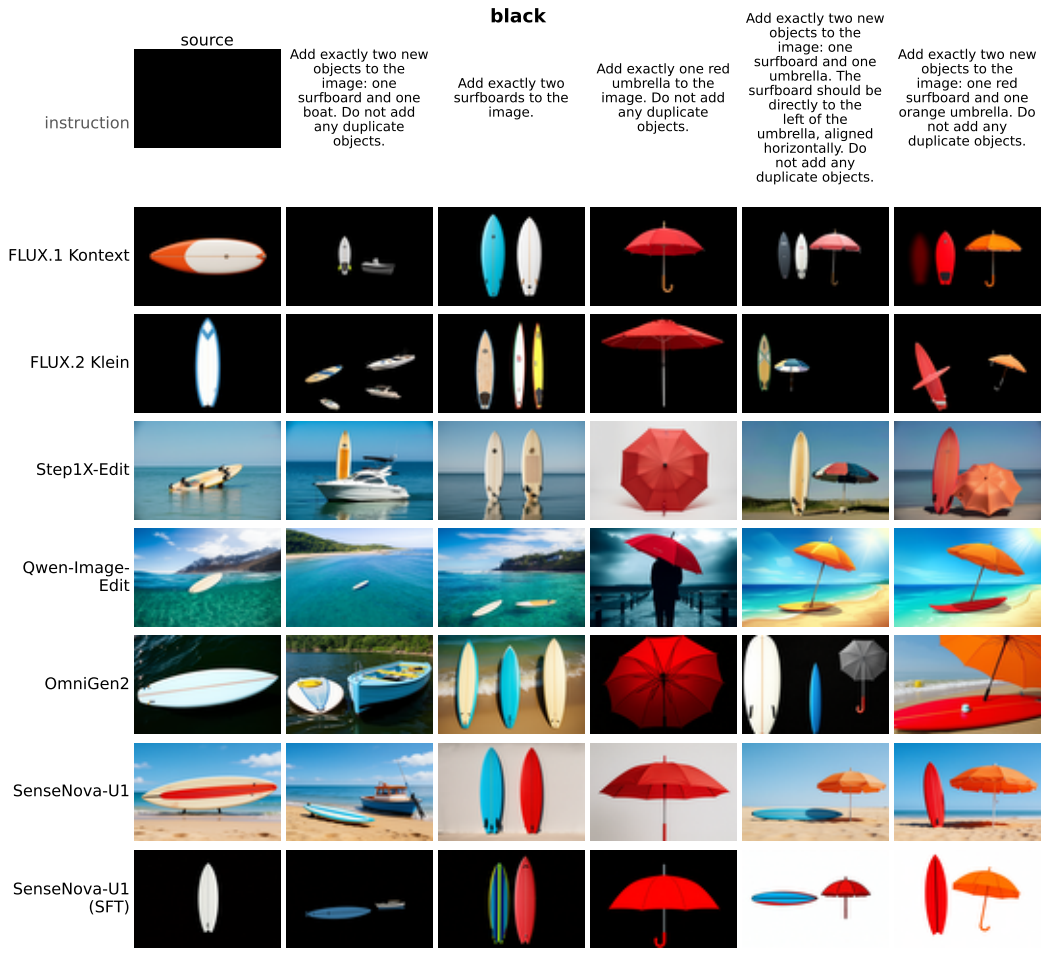


Figure 14: **Black.**

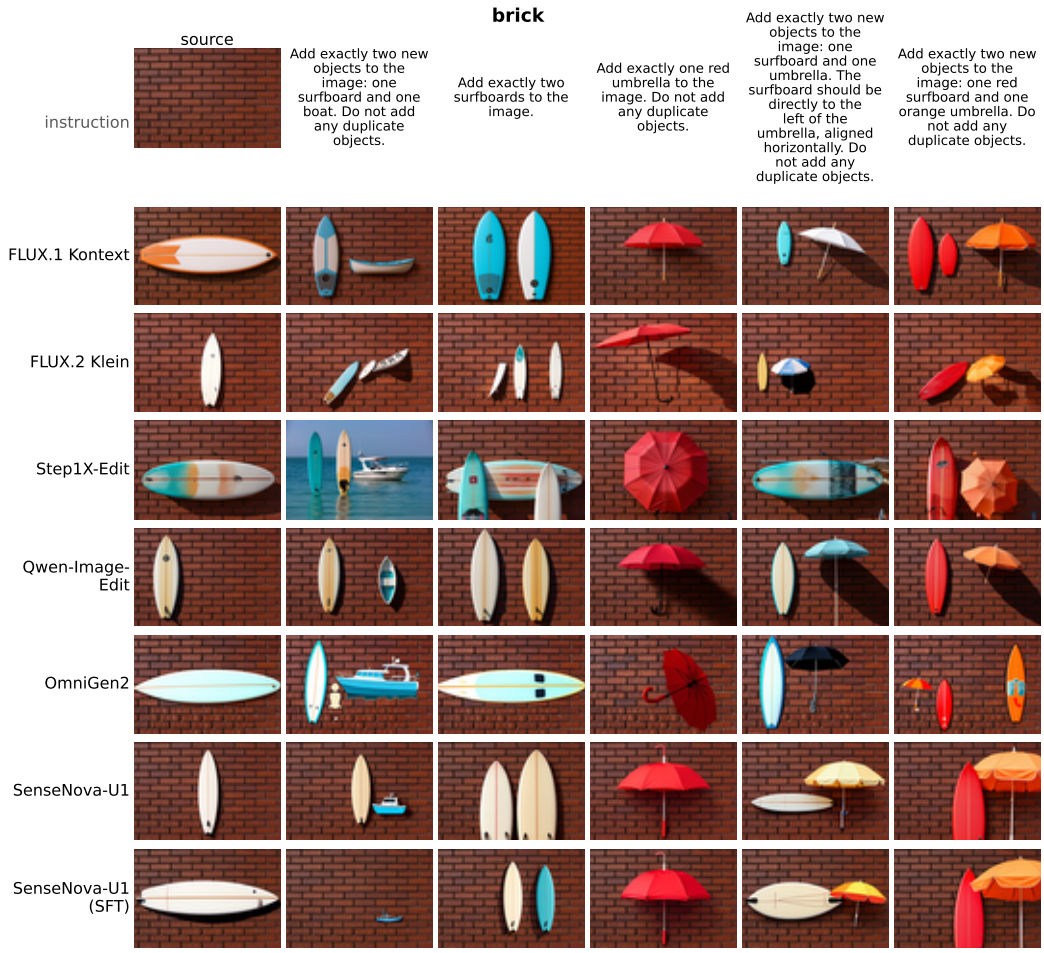


Figure 15: **Brick.**

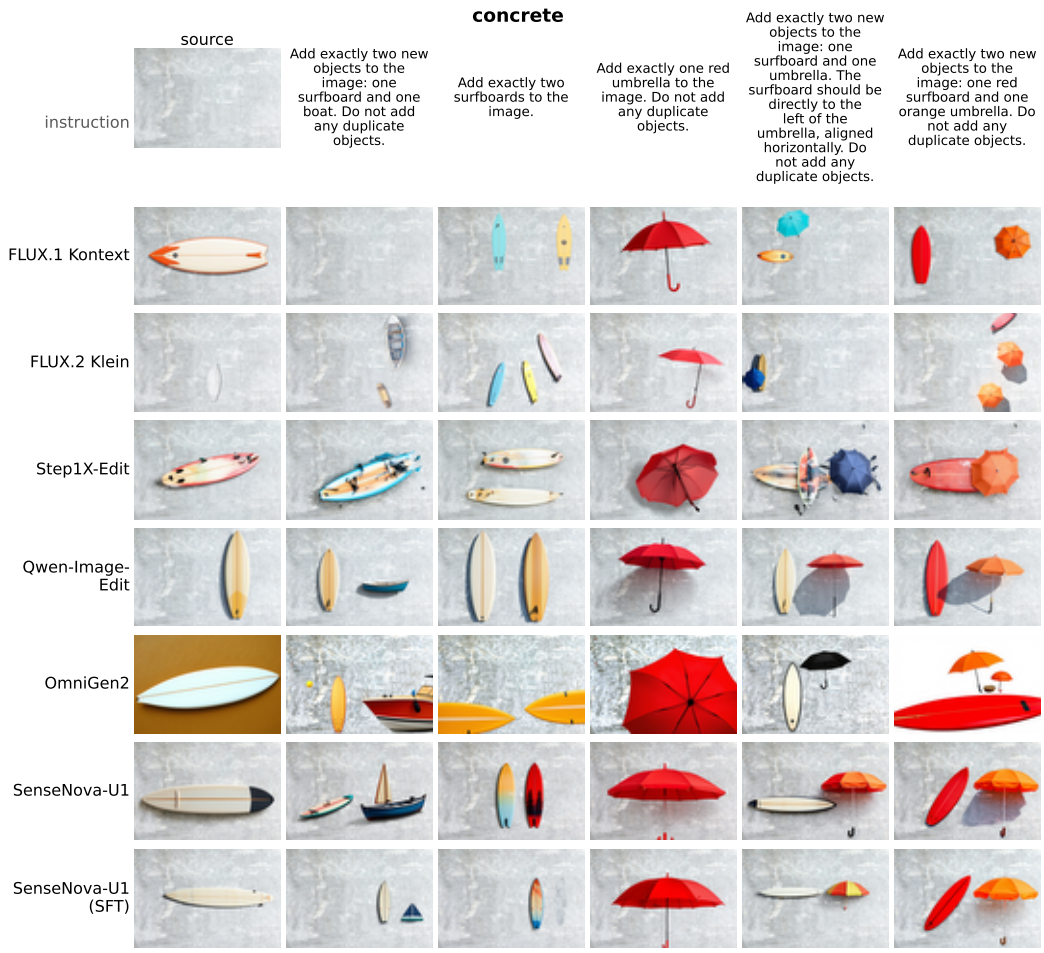


Figure 16: **Concrete.**

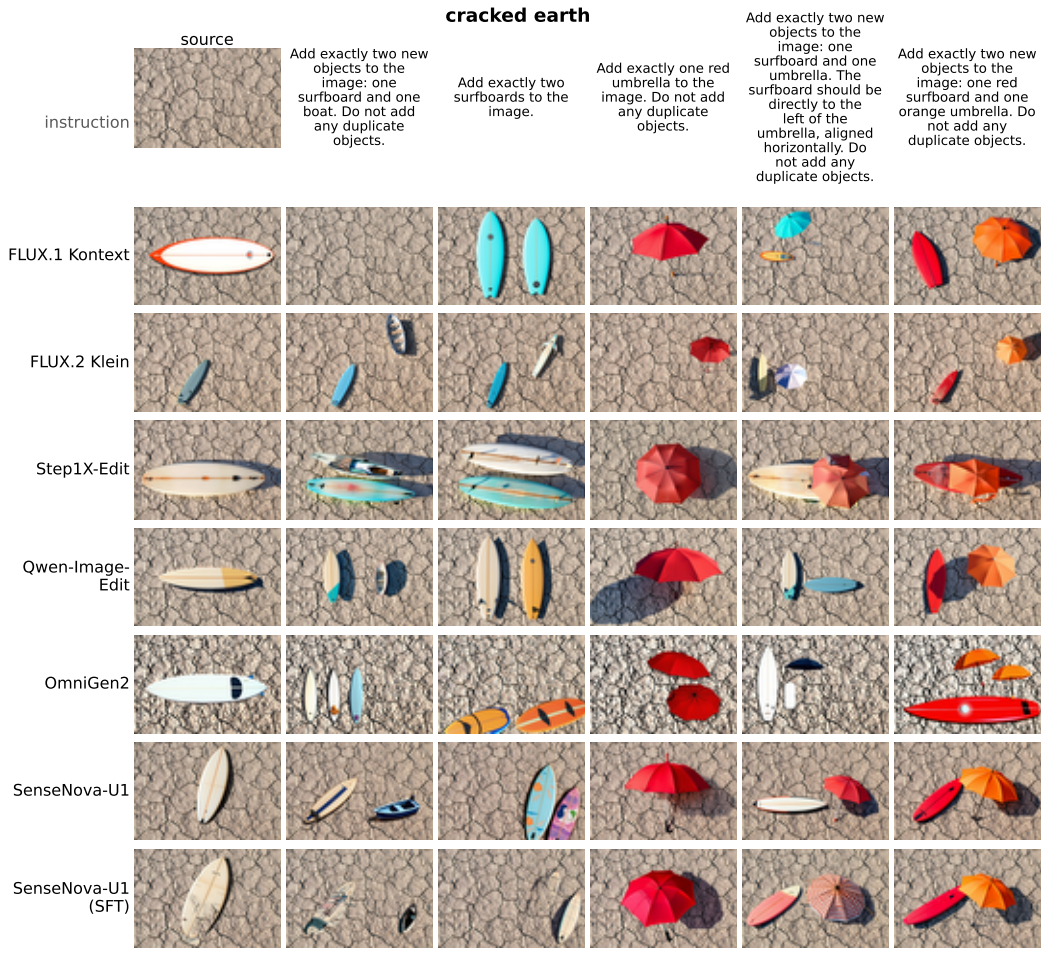


Figure 17: **Cracked earth.**

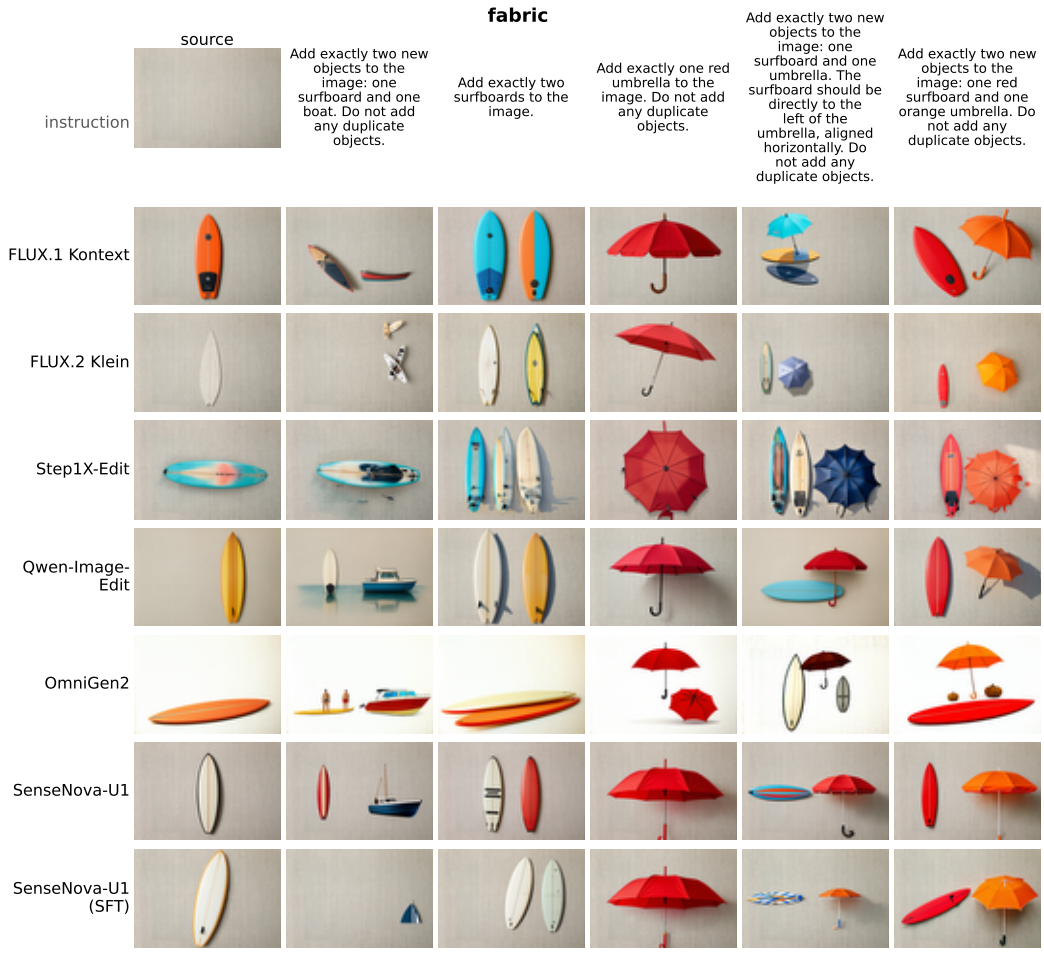


Figure 18: **Fabric.**

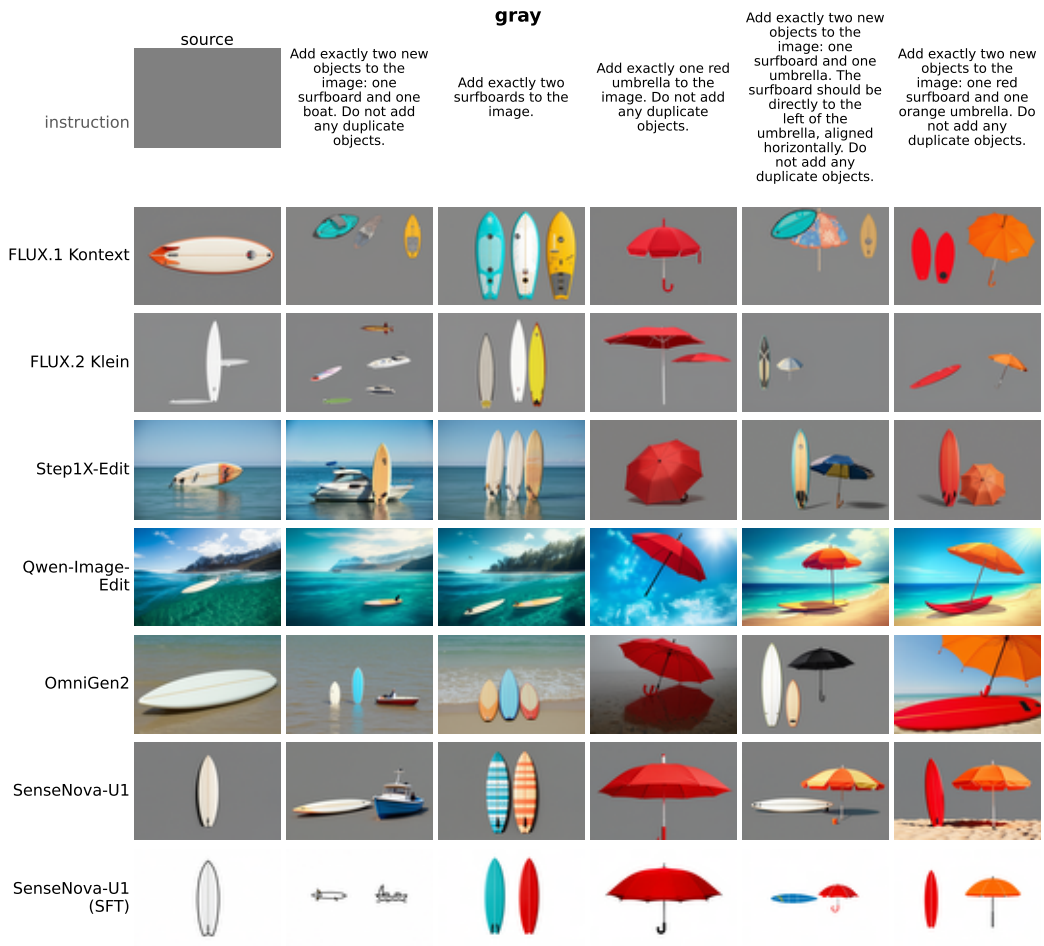


Figure 19: **Gray.**

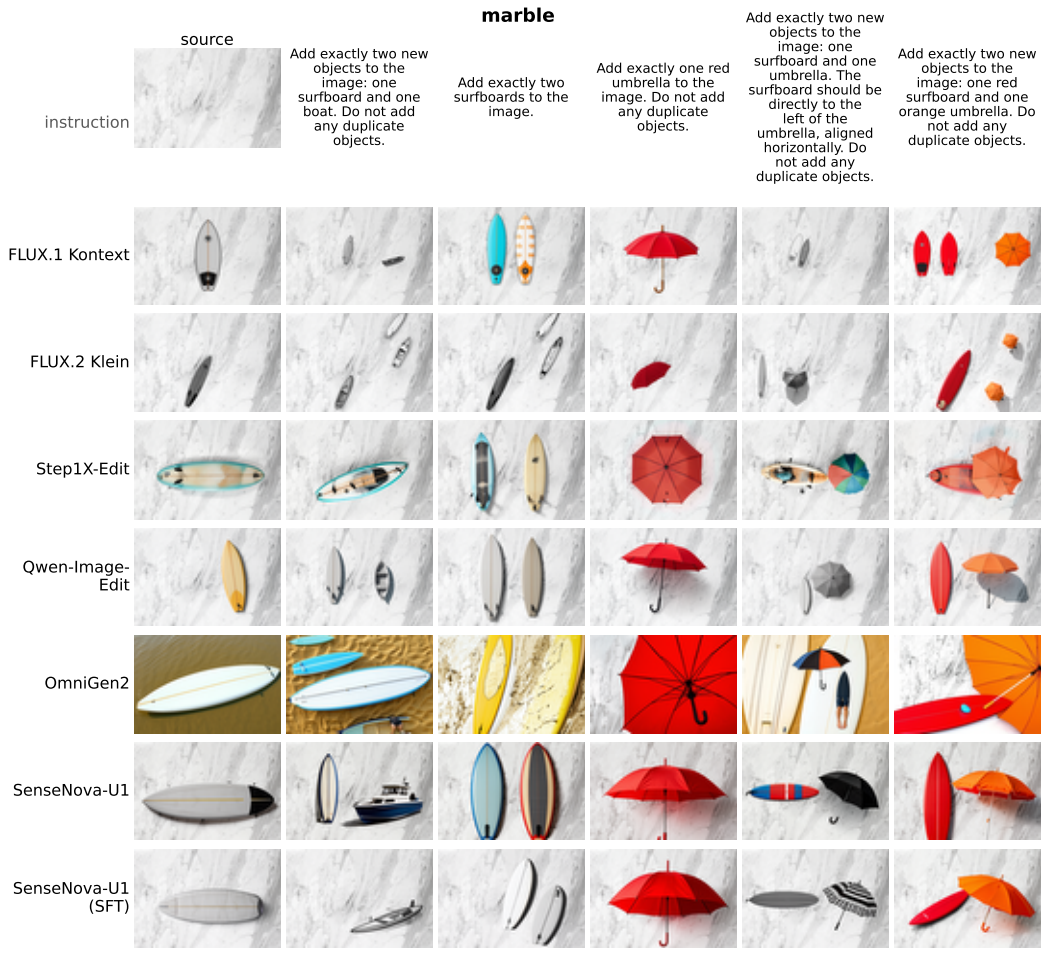


Figure 20: **Marble.**

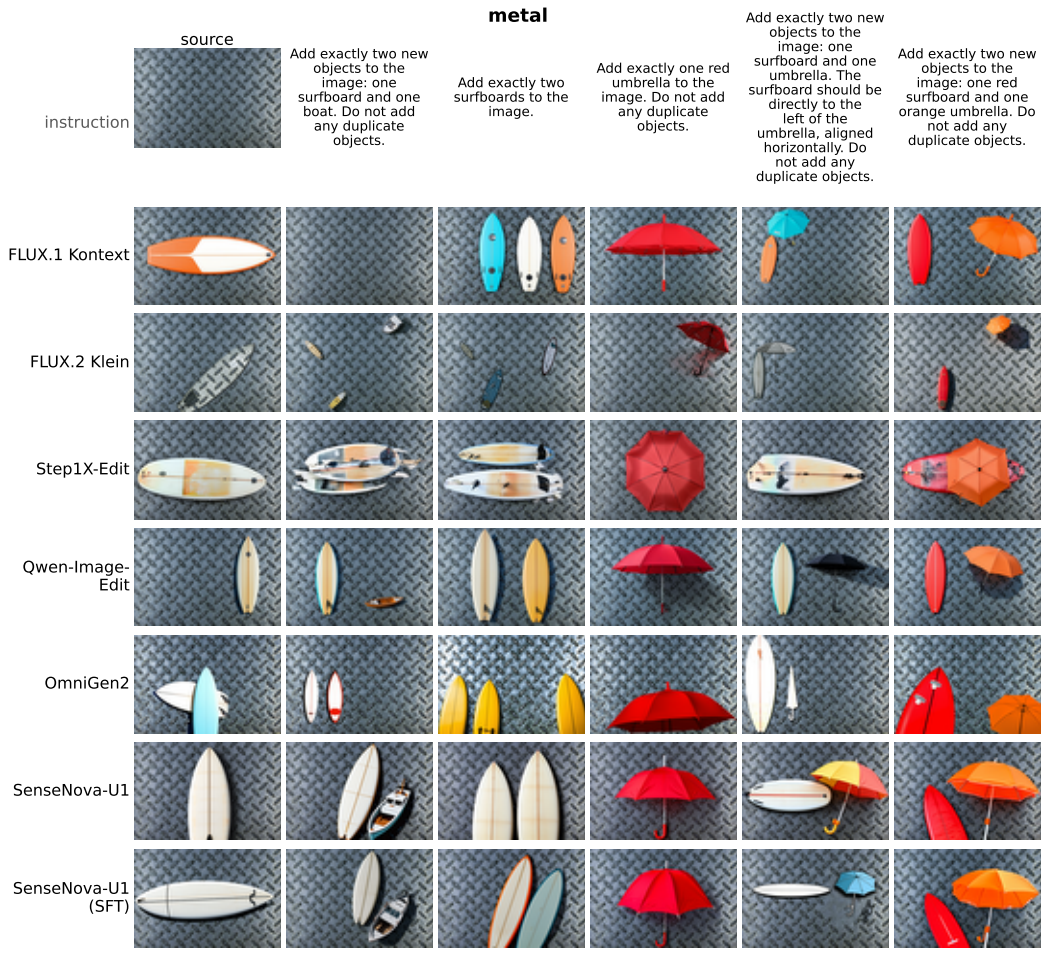


Figure 21: **Metal.**

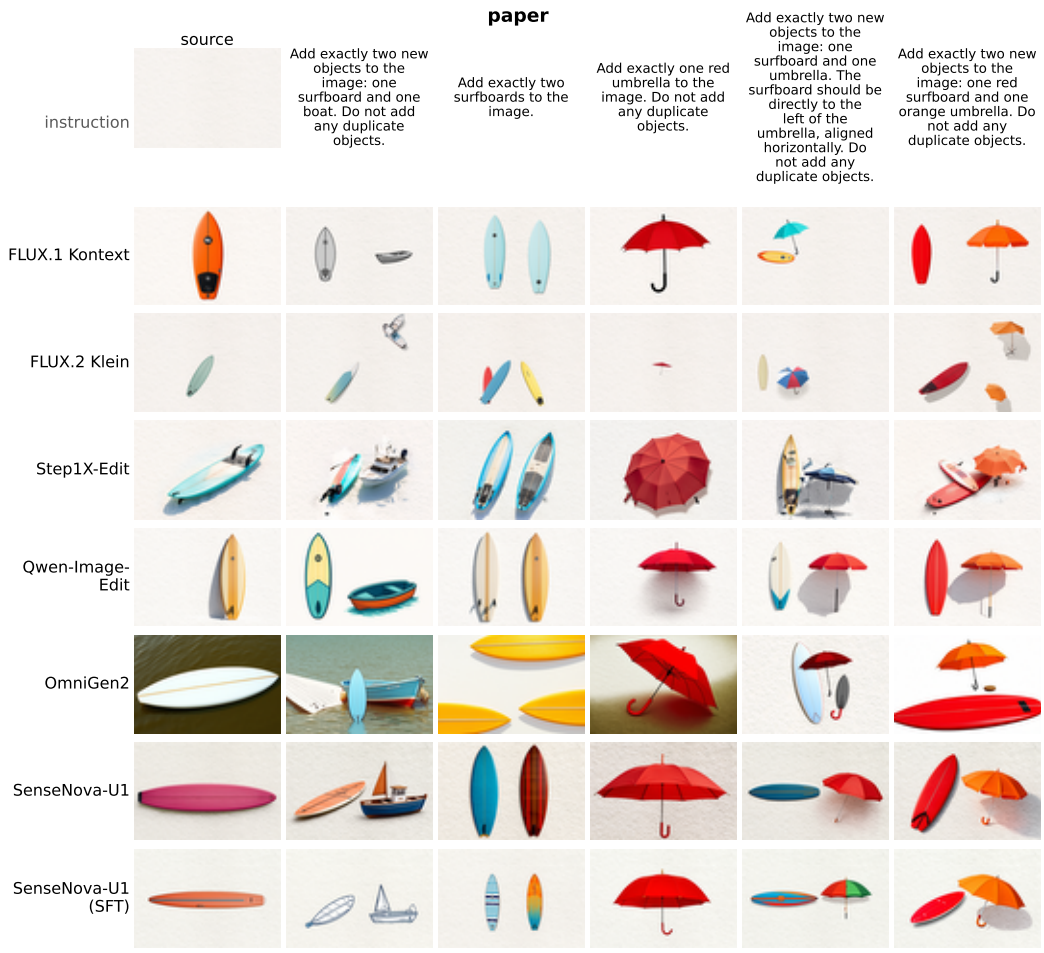


Figure 22: **Paper.**

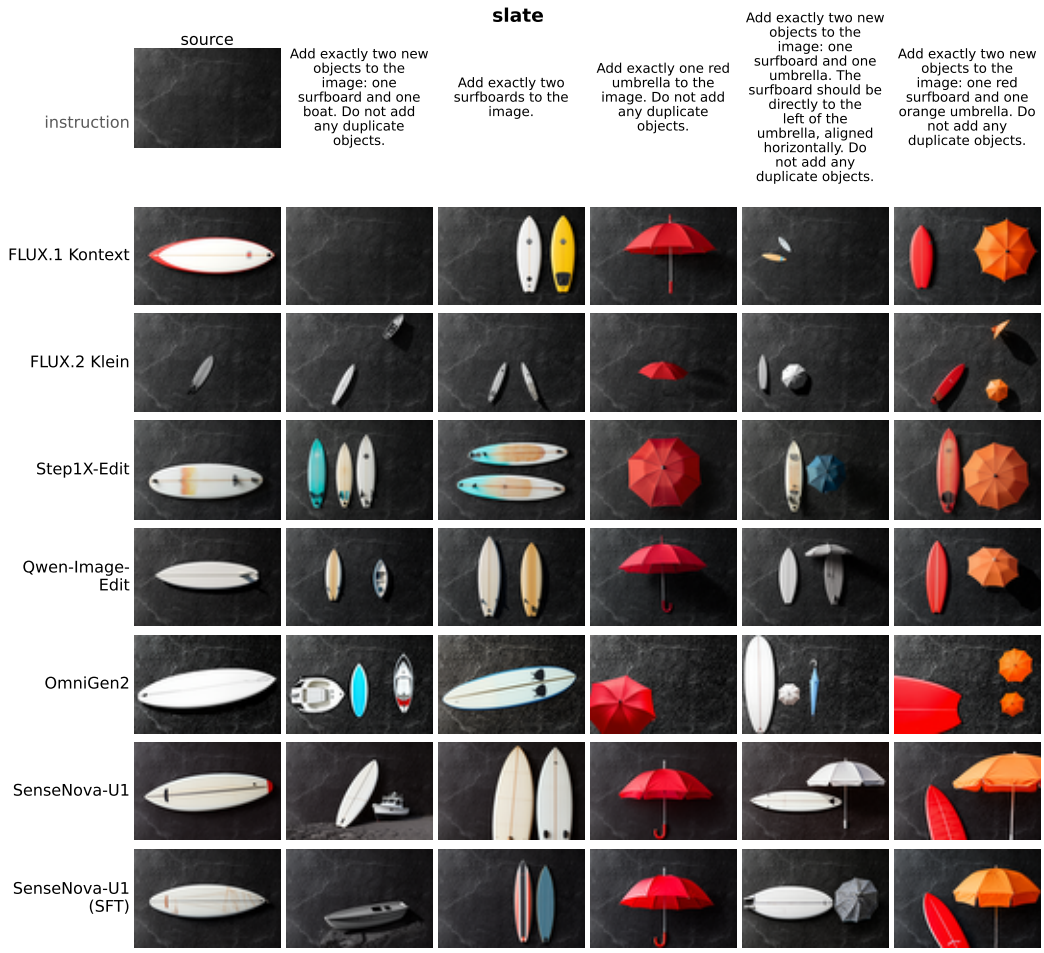


Figure 23: **Slate.**

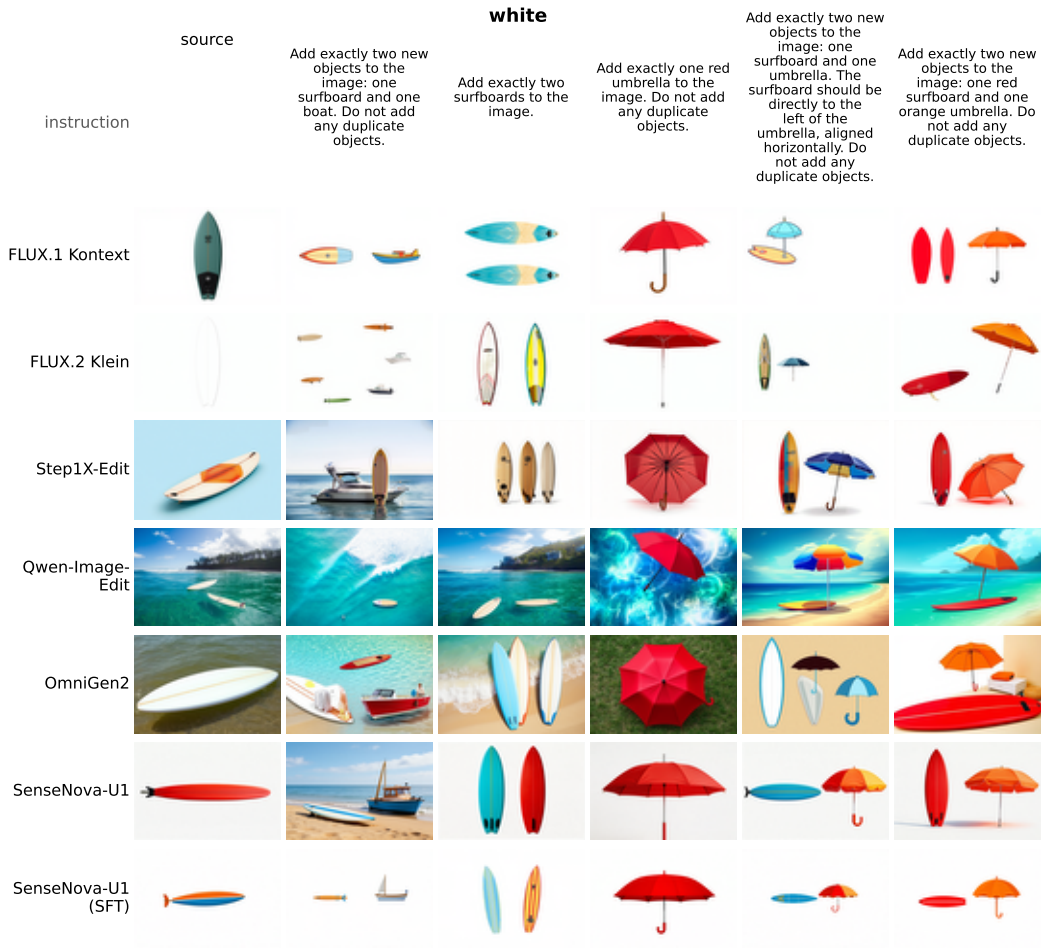


Figure 24: **White.**

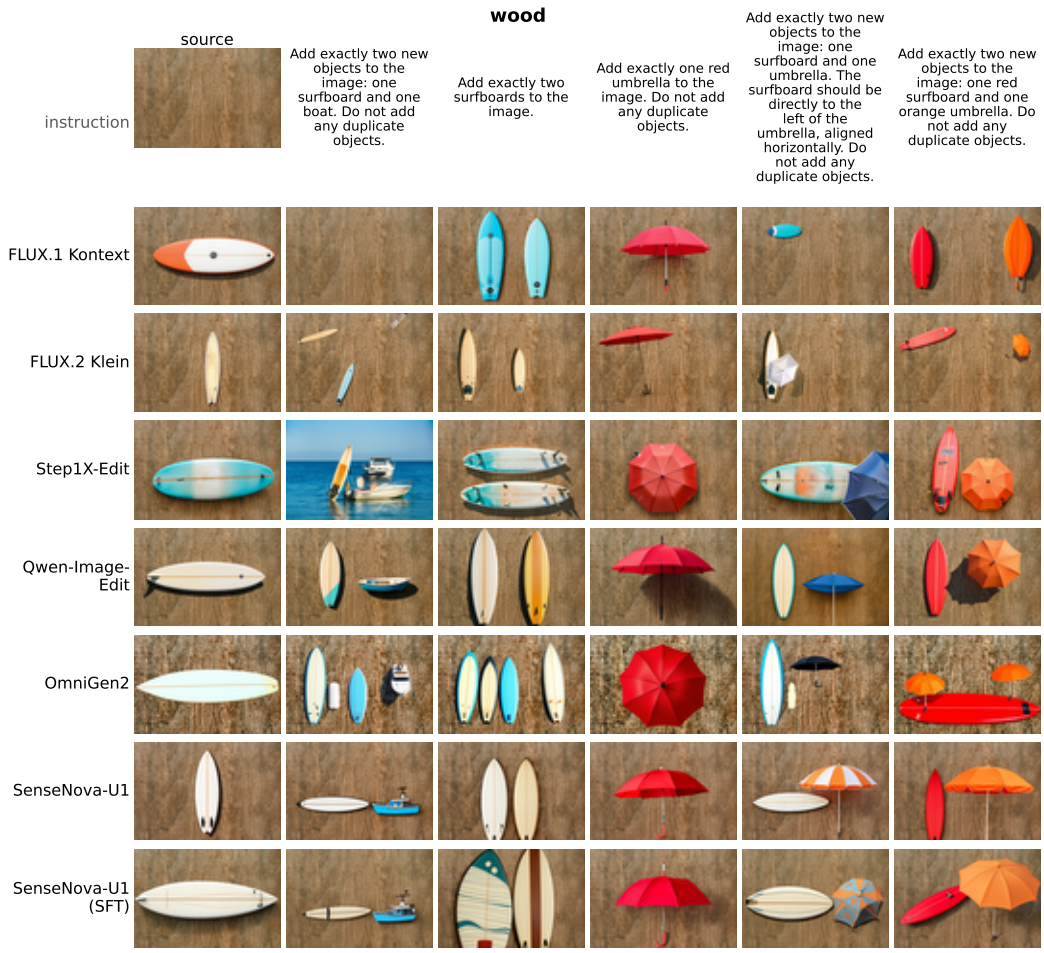


Figure 25: **Wood.**

C.4 LANDSCAPE QUALITATIVE EXAMPLES

Figures 26 to 50 show example edits on each of the Landscape for all models. Each figure shows the source image and the outputs of each model for the same prompts.

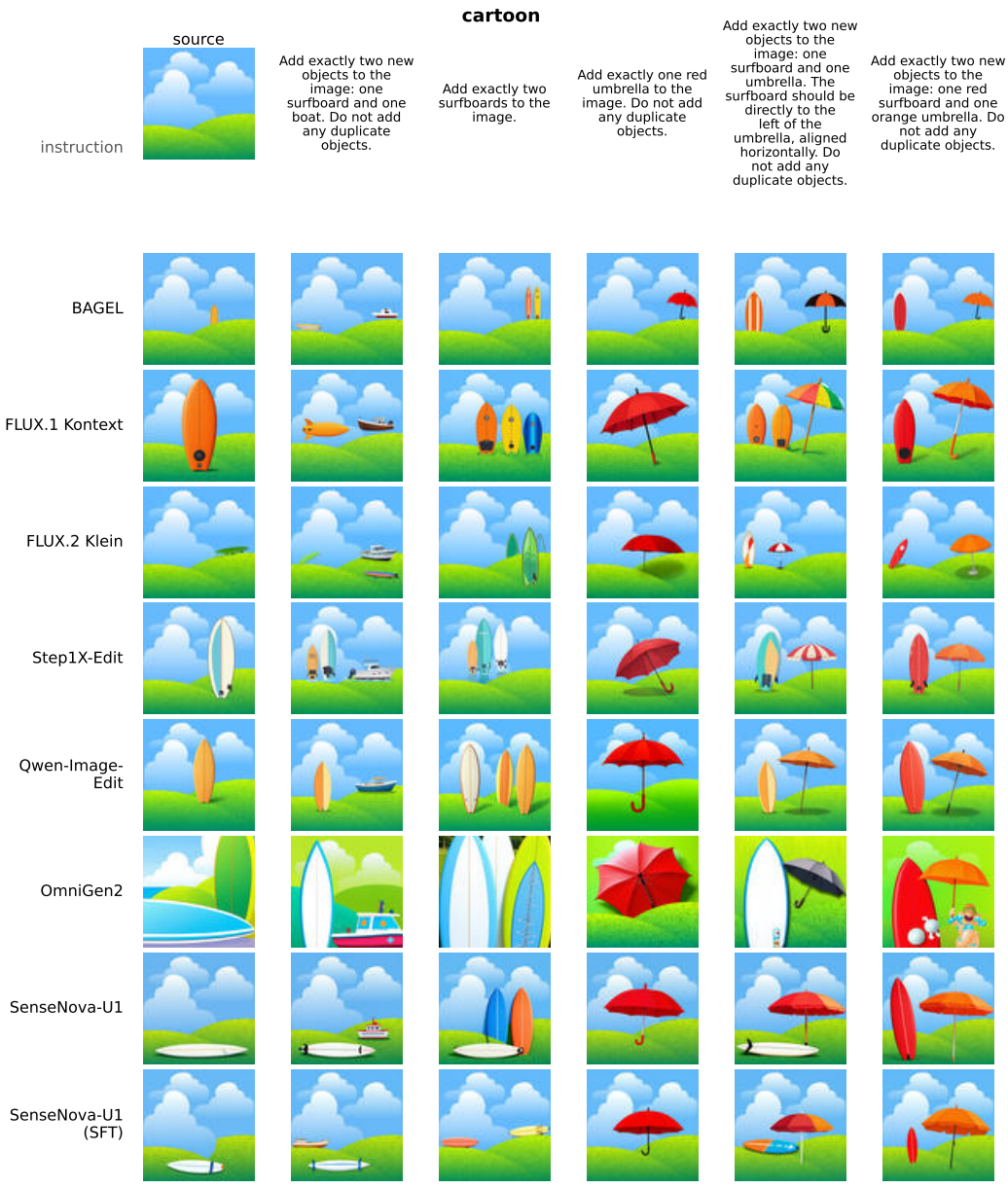


Figure 26: **Cartoon.**

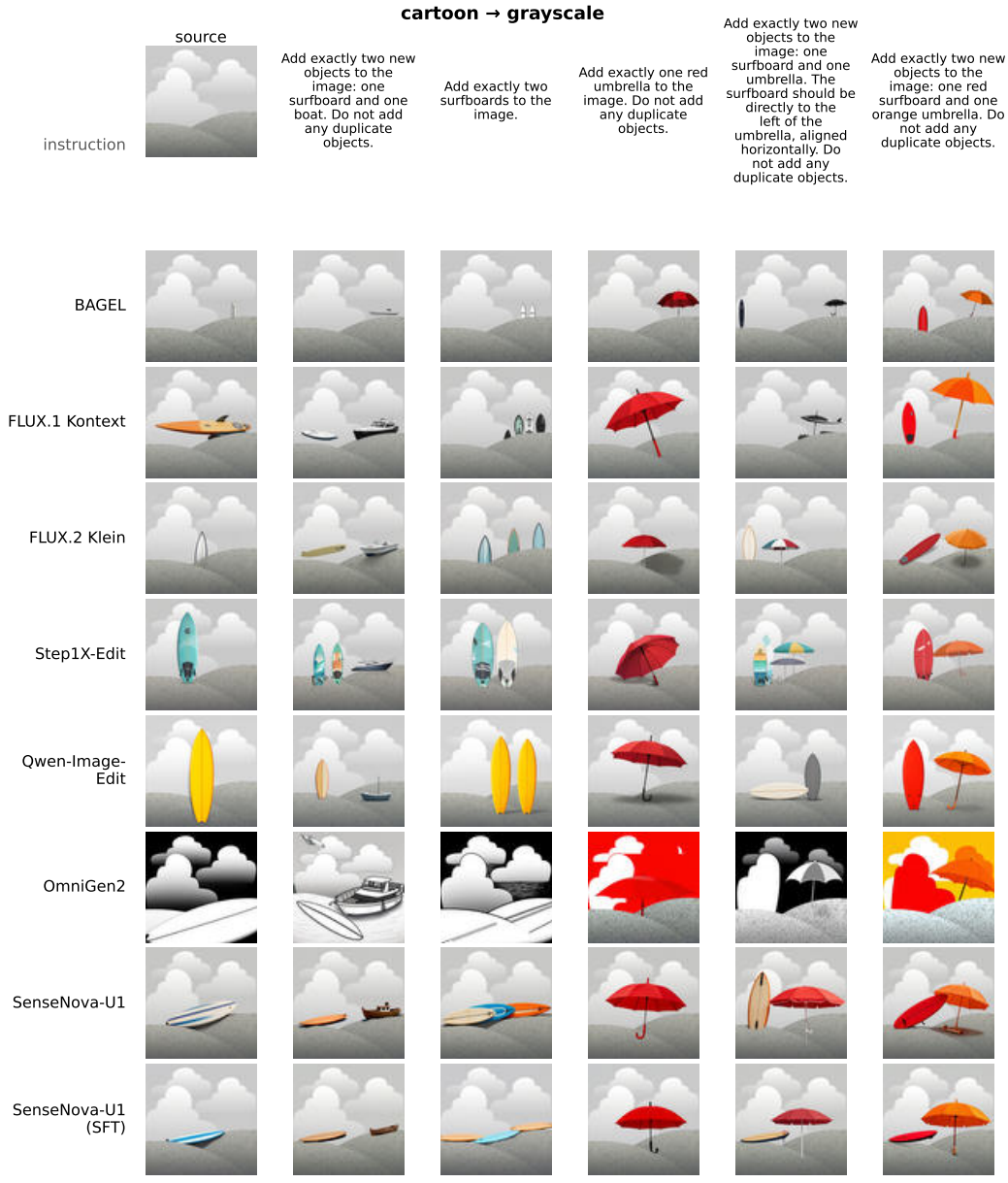


Figure 27: **Cartoon → grayscale.**

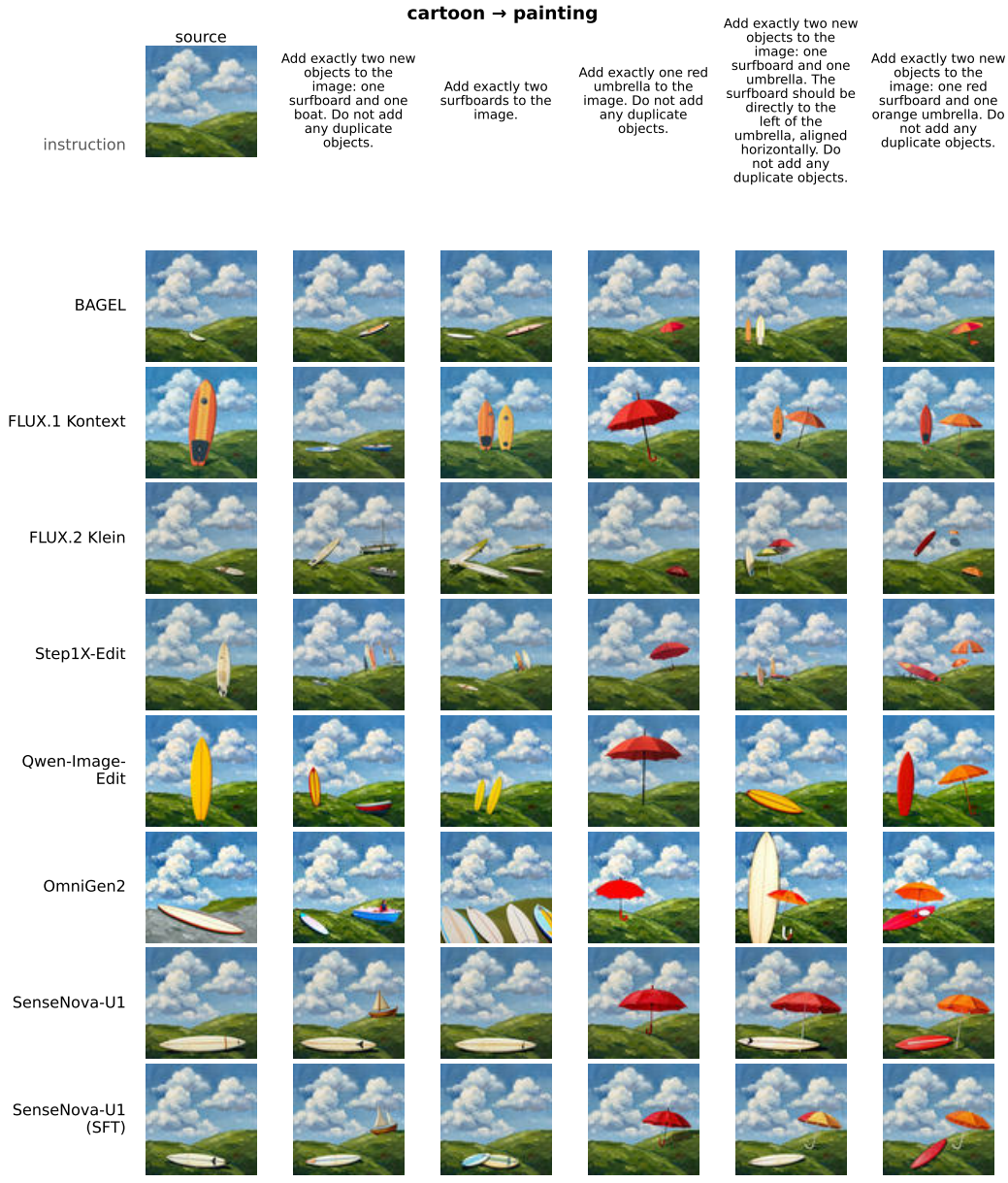


Figure 28: **Cartoon → painting.**



Figure 29: **Cartoon → Photo.**

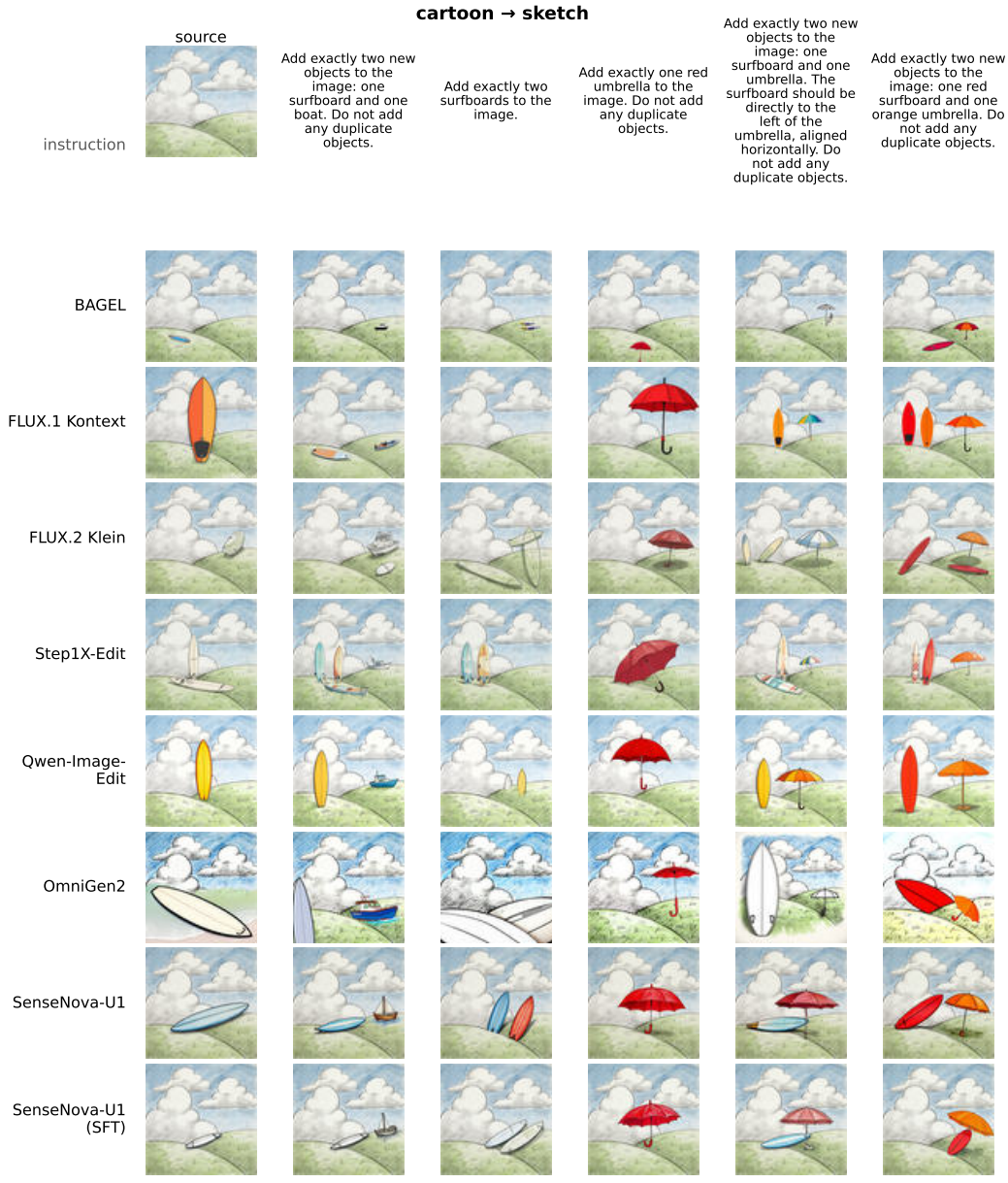


Figure 30: **Cartoon → sketch.**

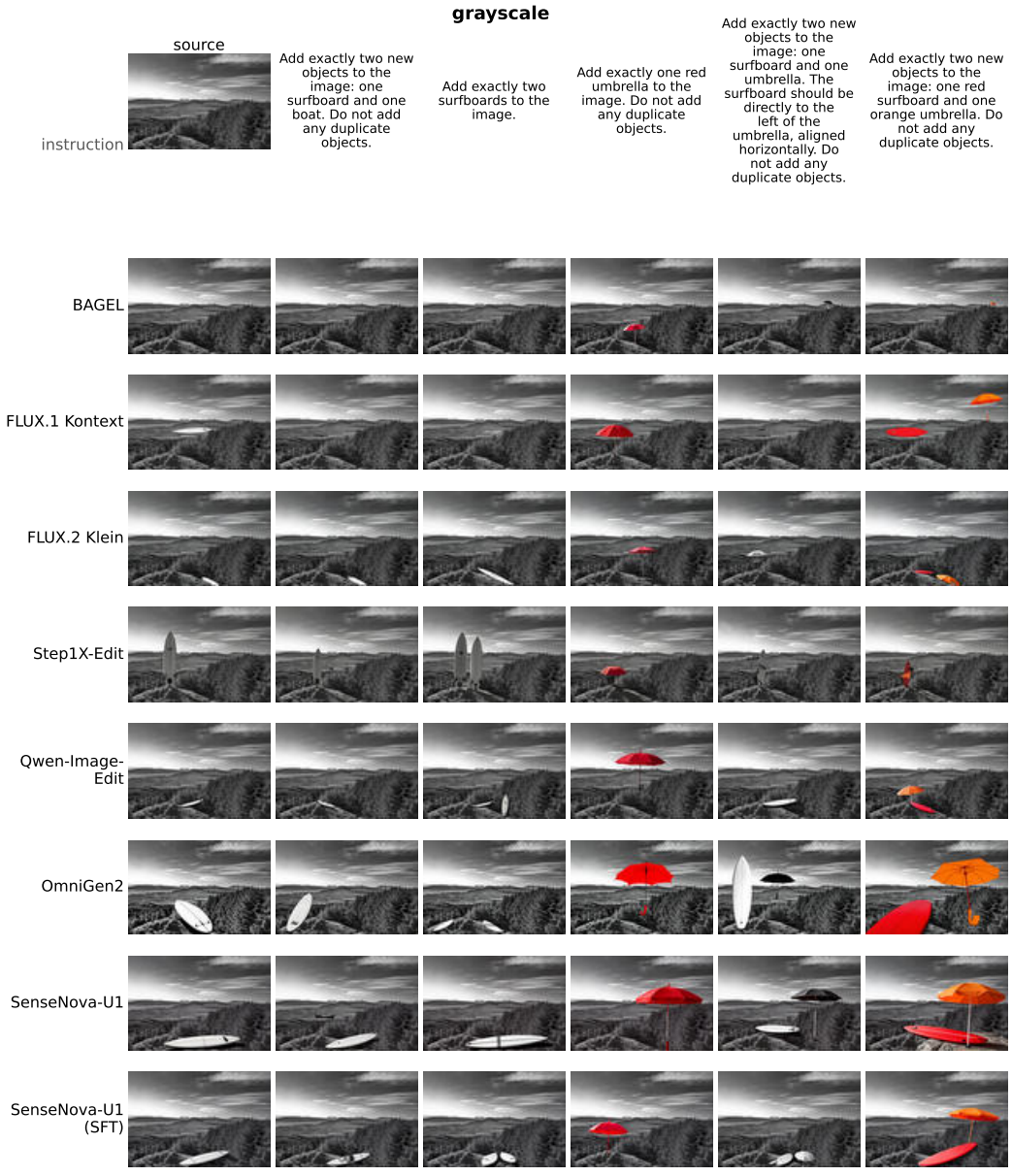


Figure 31: **Grayscale.**

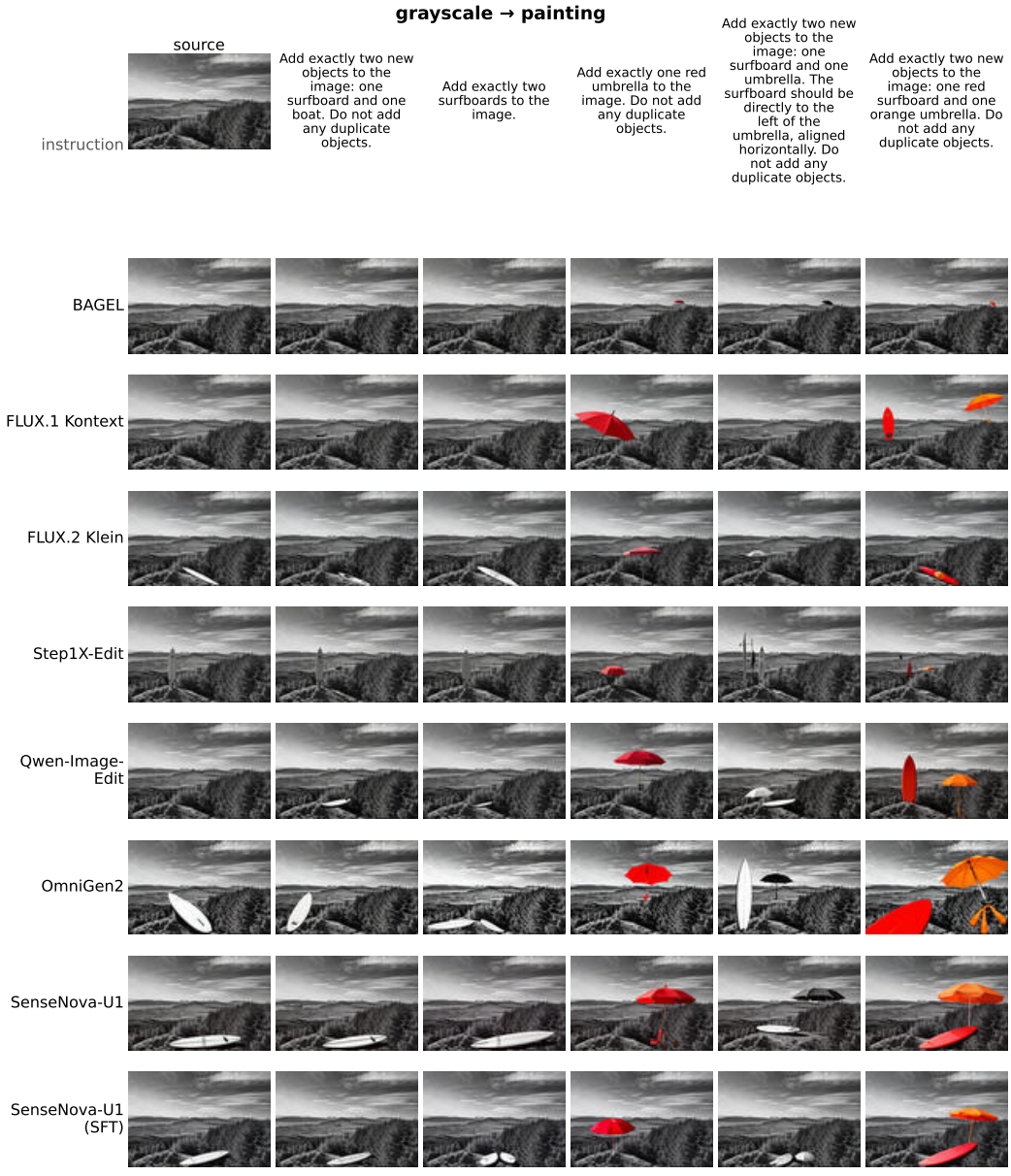


Figure 32: **Grayscale → painting.**

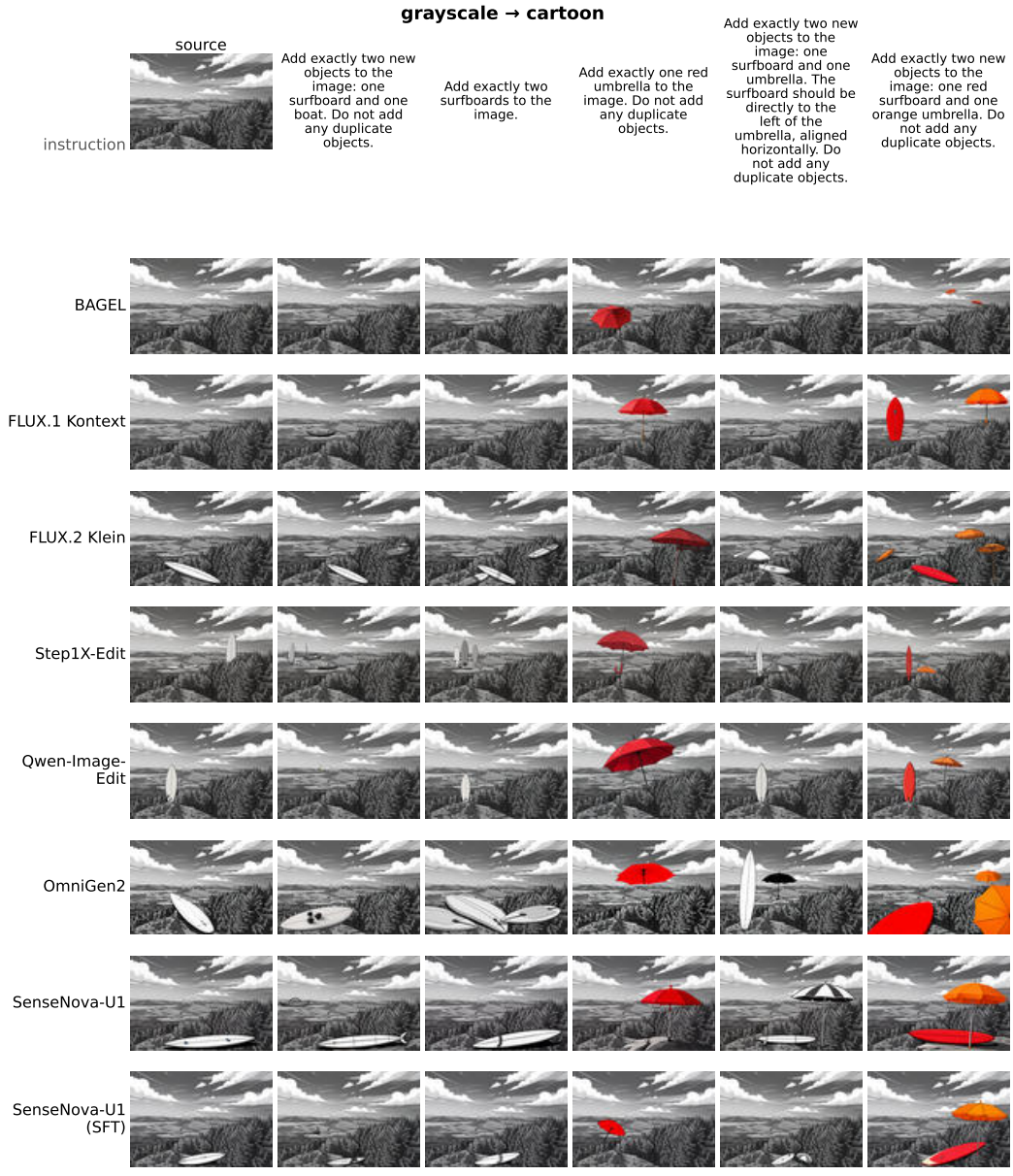


Figure 33: **Grayscale → cartoon.**



Figure 34: **Grayscale → Photo.**

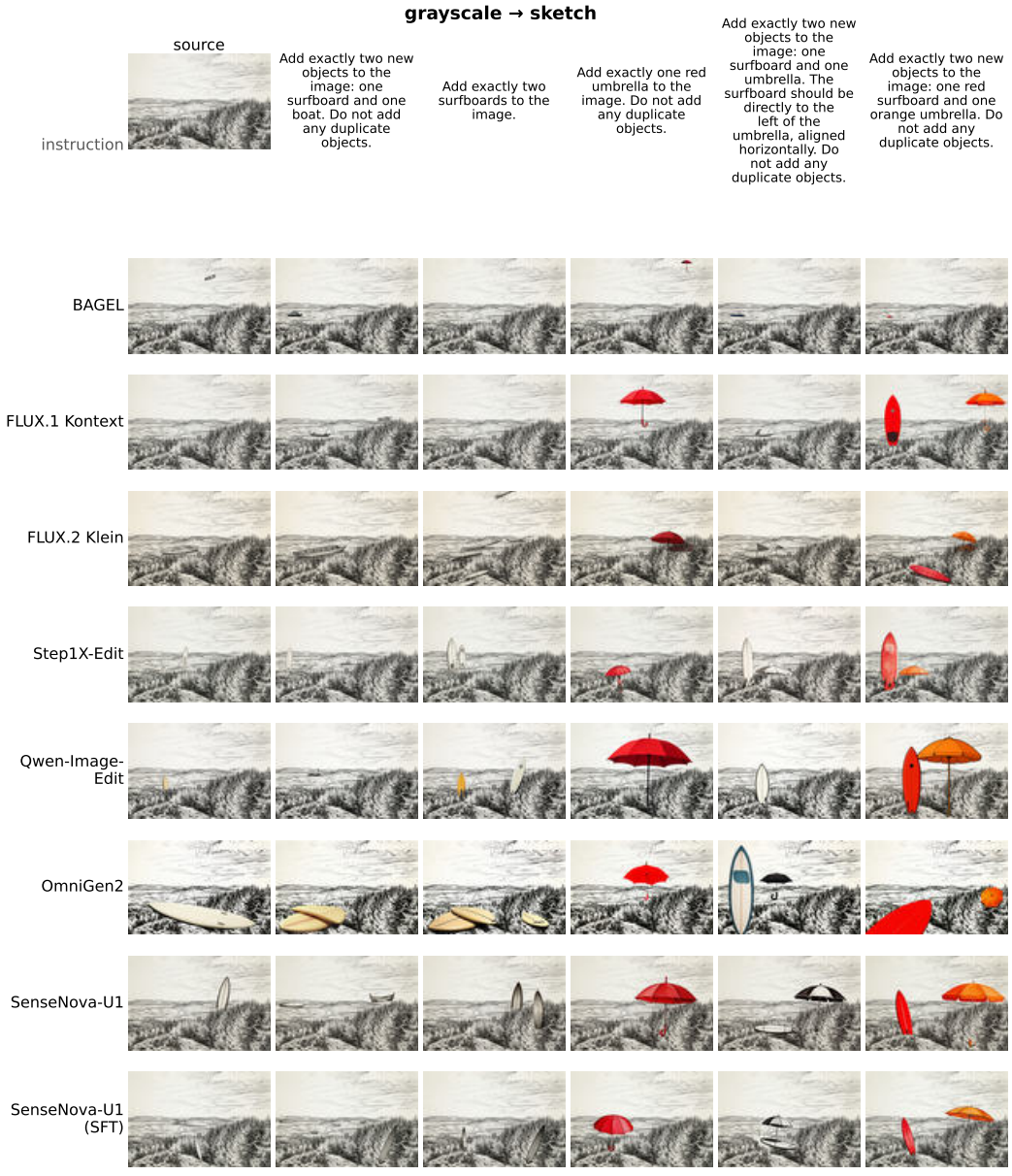


Figure 35: **Grayscale → sketch.**



Figure 36: **Painting.**



Figure 37: **Painting → cartoon.**

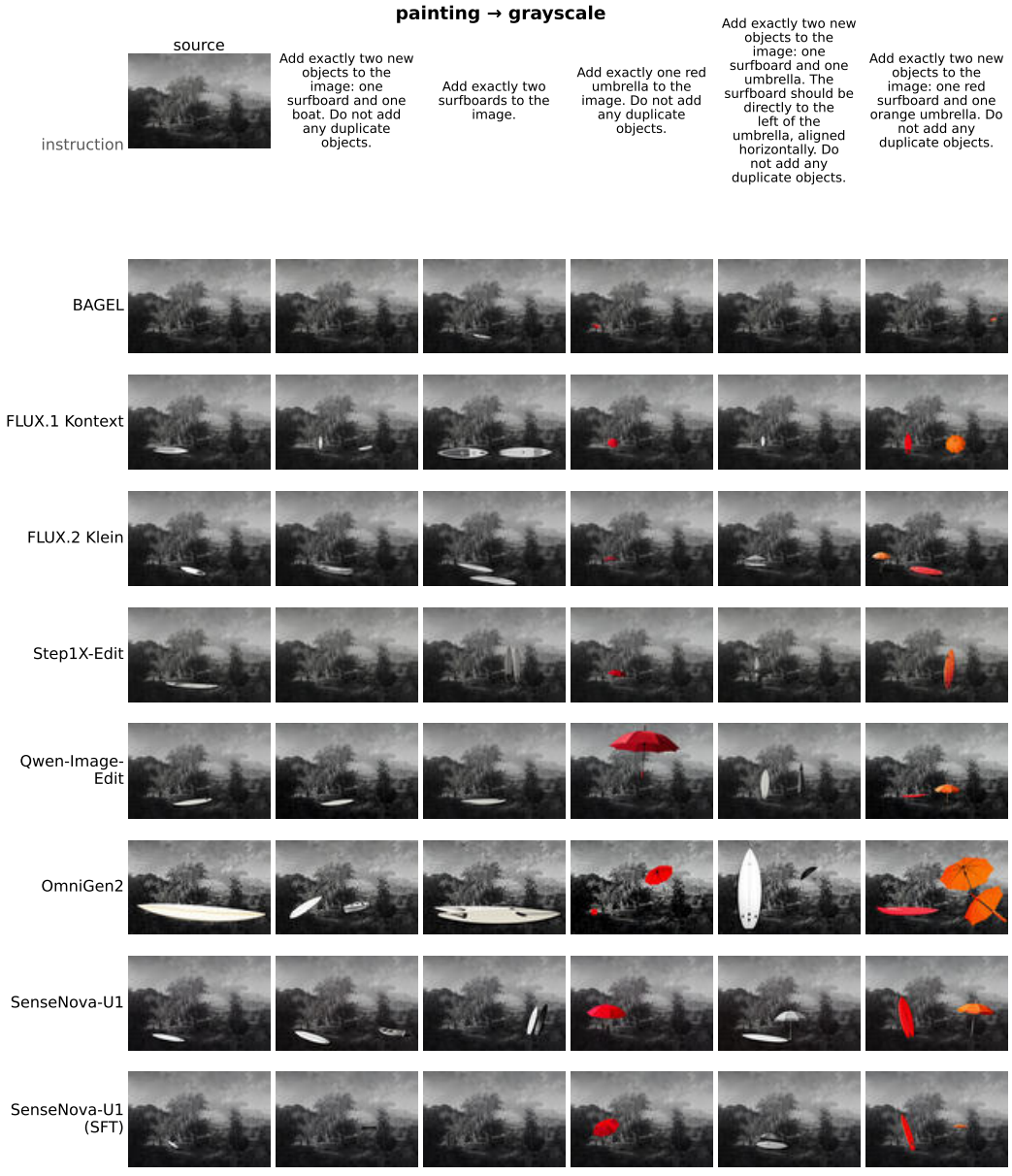


Figure 38: **Painting → grayscale.**

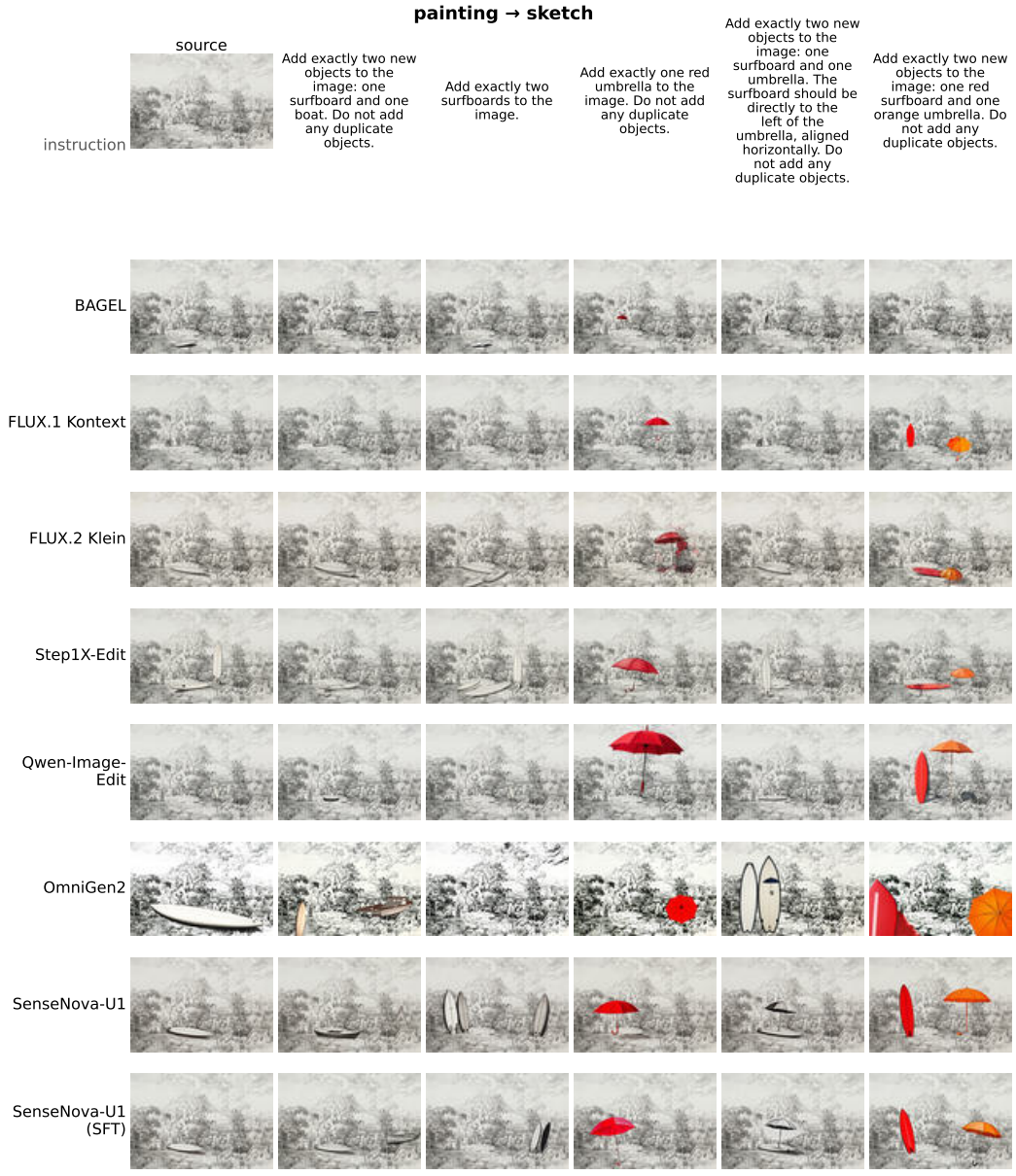


Figure 39: **Painting → sketch.**



Figure 40: **Painting → photo.**



Figure 41: **Real.**

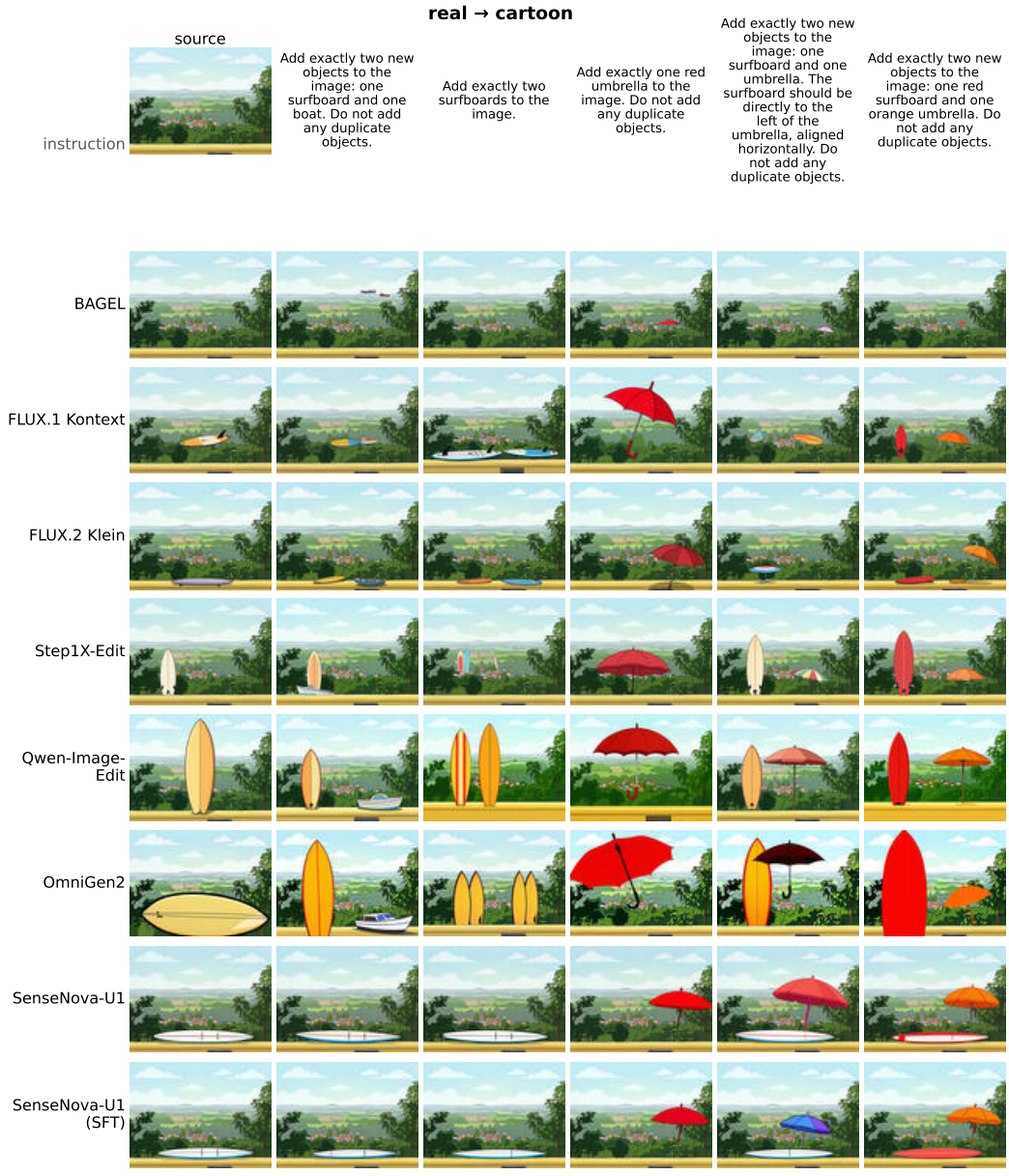


Figure 42: **Real → cartoon.**

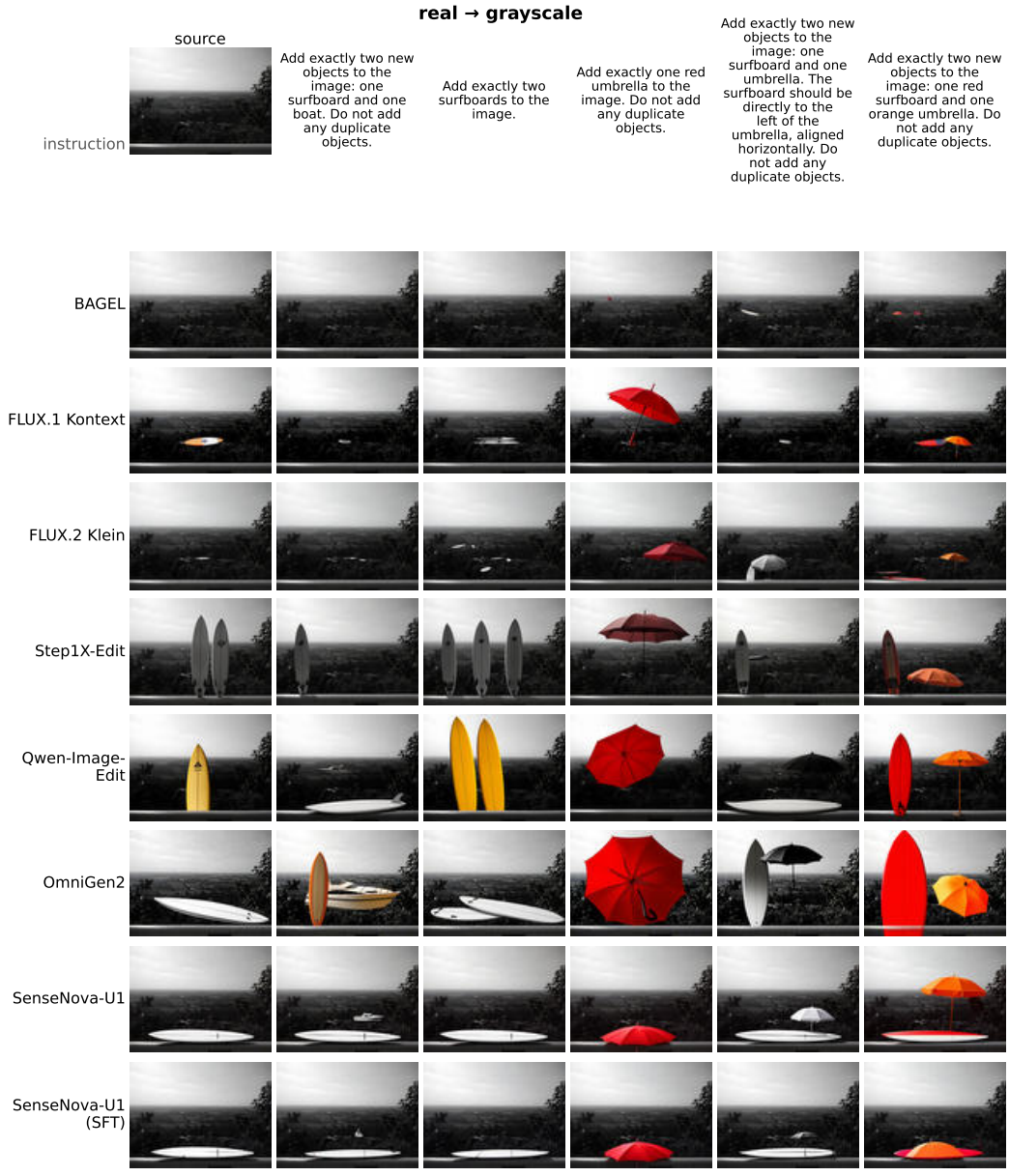


Figure 43: **Real → grayscale.**

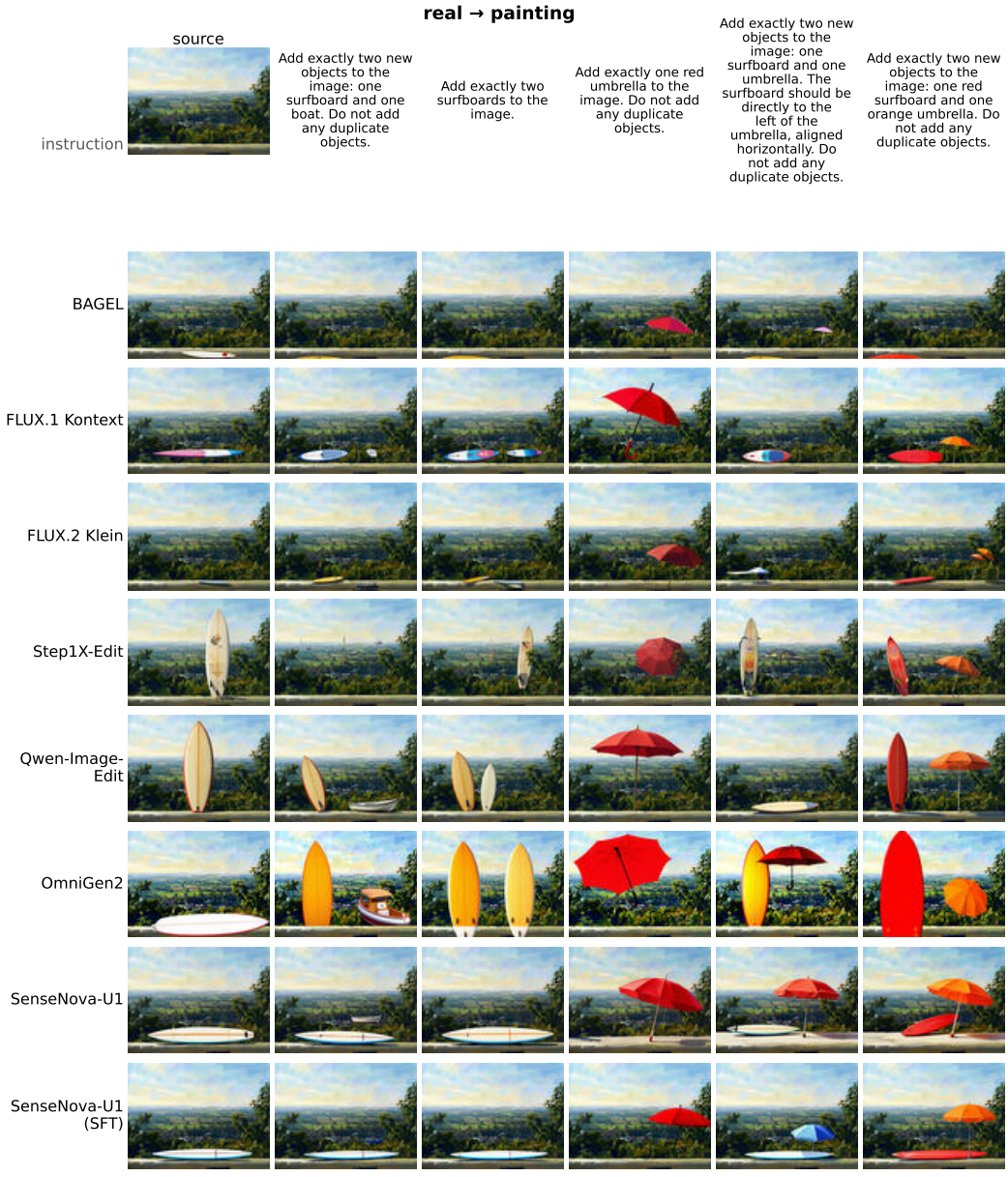


Figure 44: **Real → painting.**

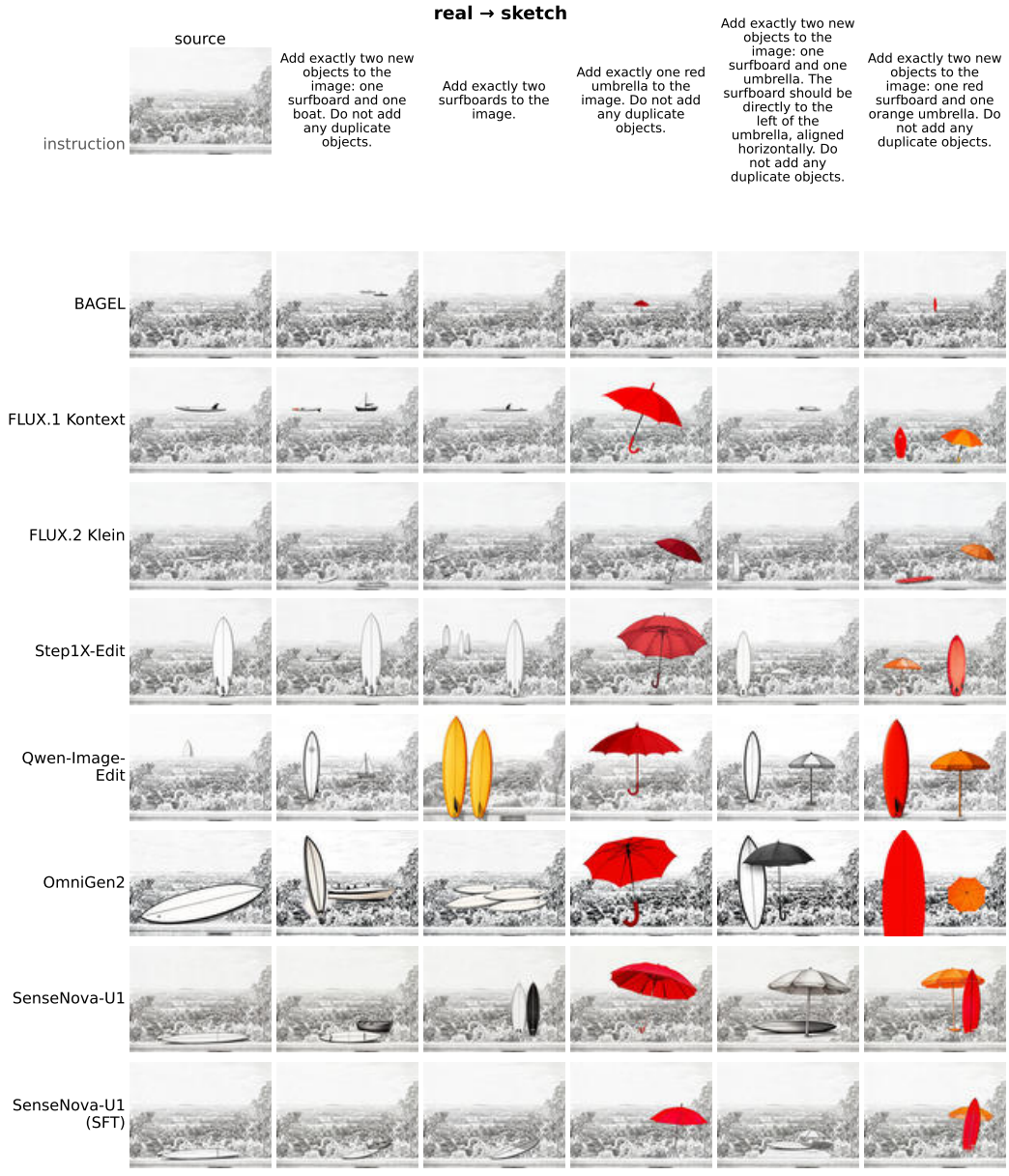


Figure 45: **Real → sketch.**

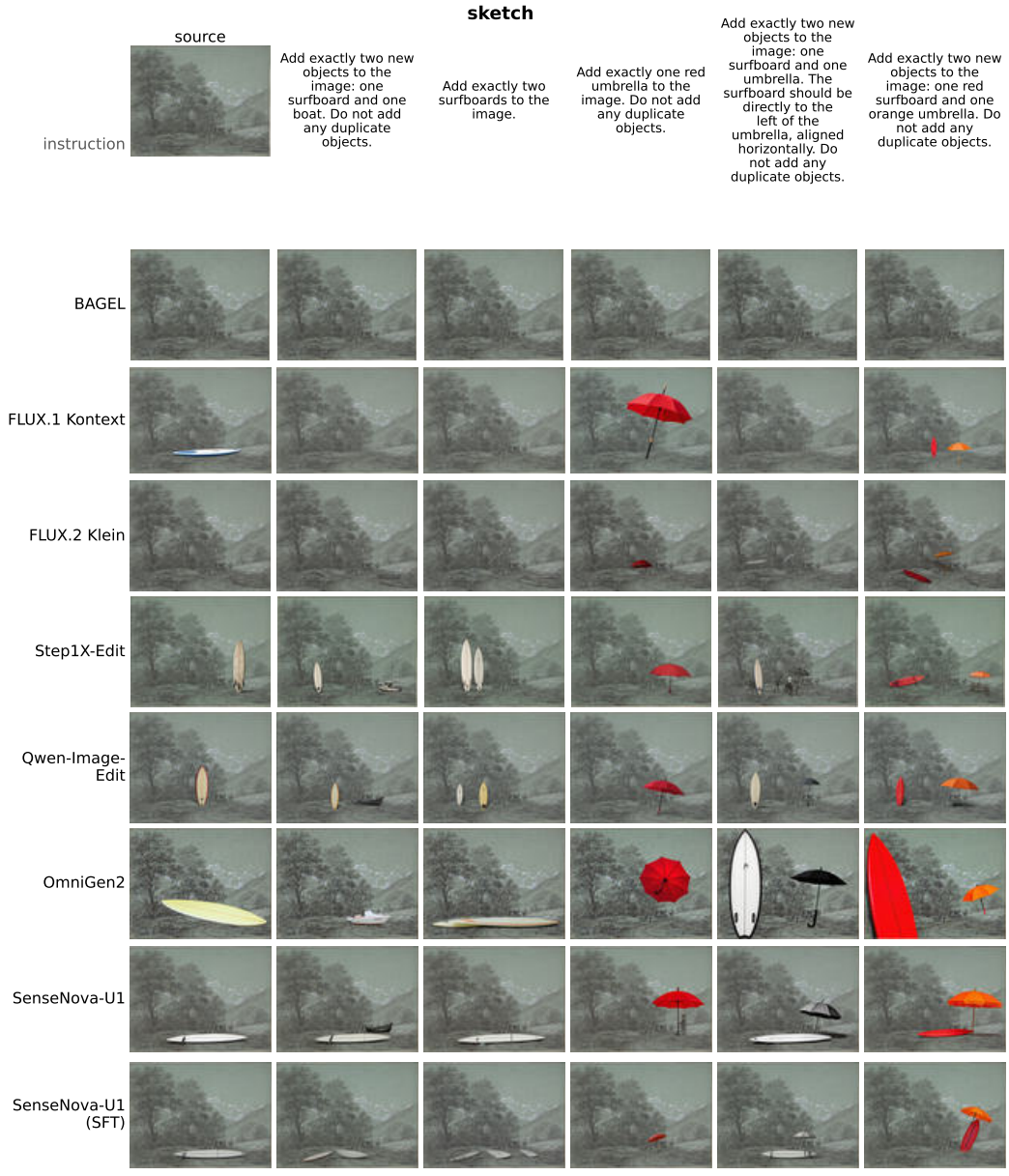


Figure 46: **Sketch**

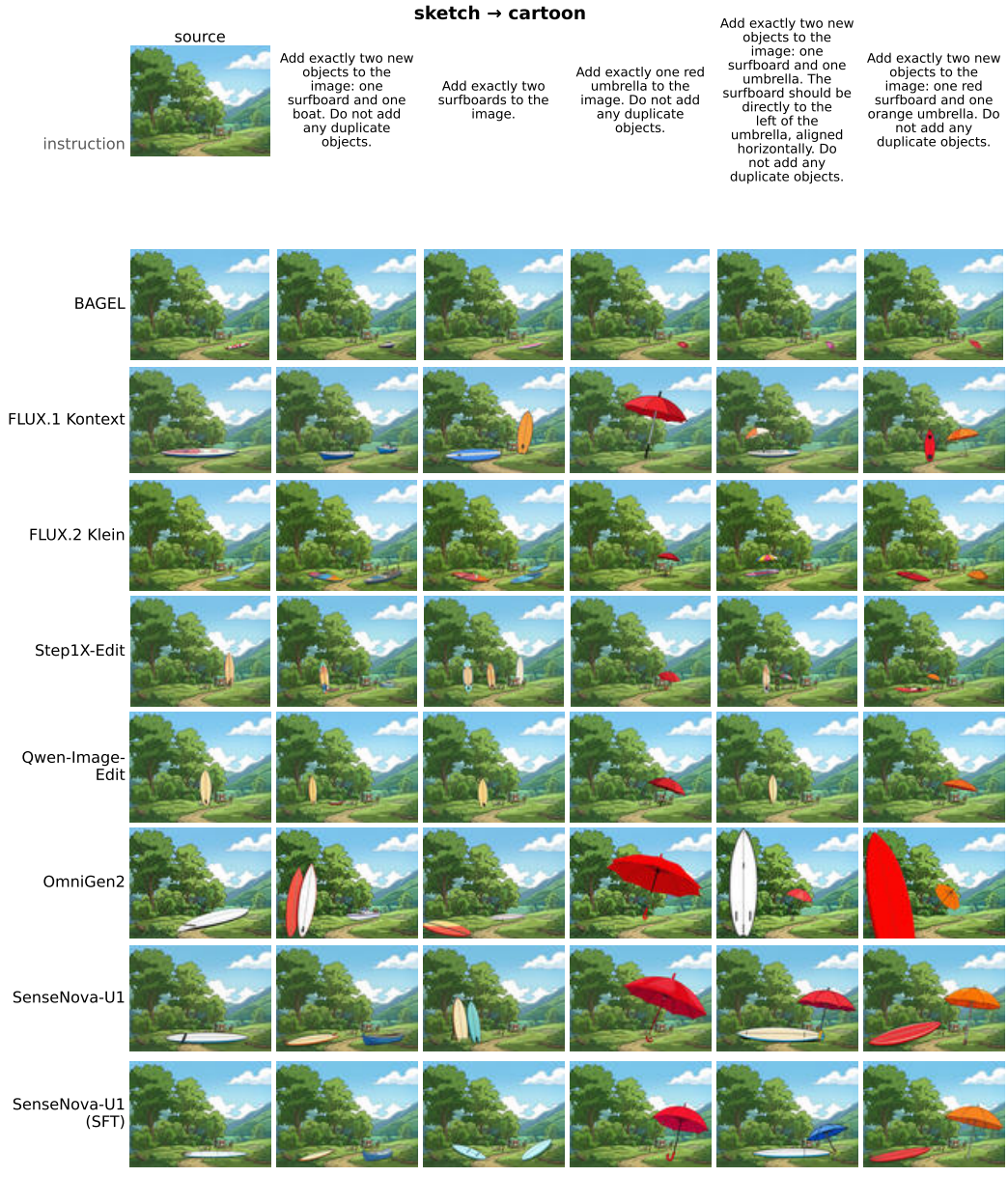


Figure 47: **Sketch → cartoon.**

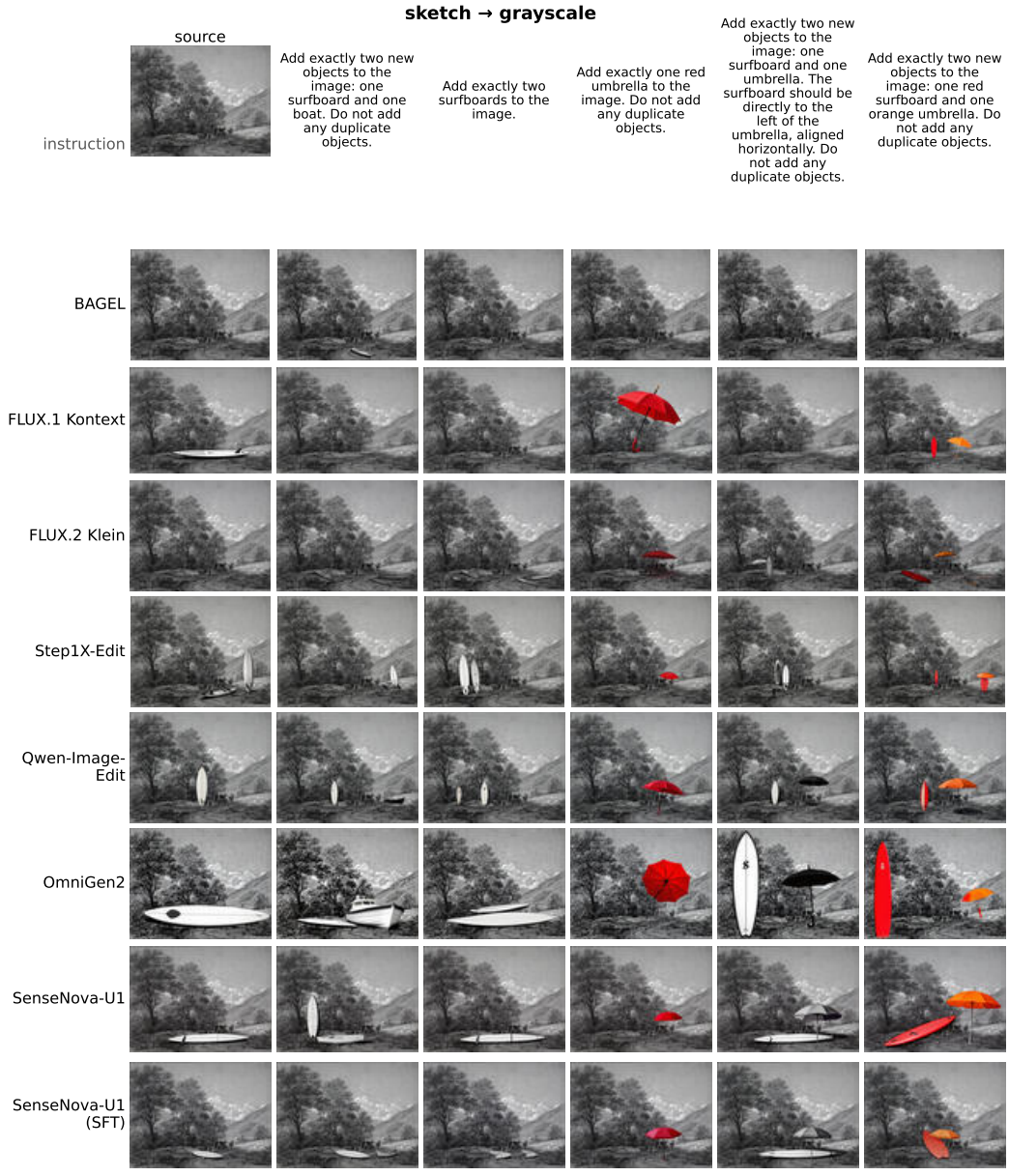


Figure 48: **Sketch → grayscale.**

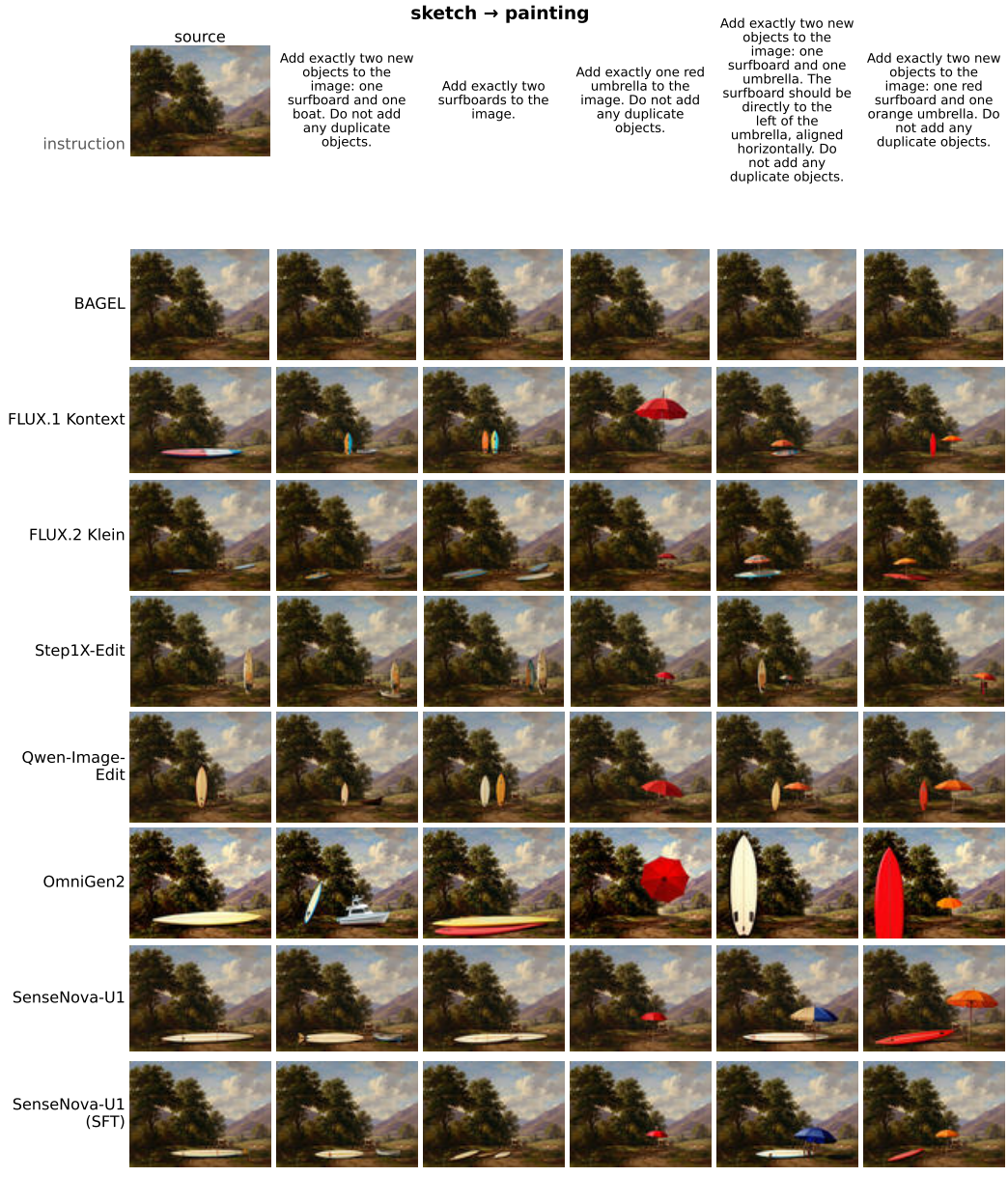


Figure 49: **Sketch → painting.**



Figure 50: **Sketch → photo.**

C.5 SCENE QUALITATIVE EXAMPLES

Figures 51 to 55 show example edits on each of the Scene for all models. Each figure shows the source image and the outputs of each model for the same prompts.

subject add: "Add a face mask to the chef's face."

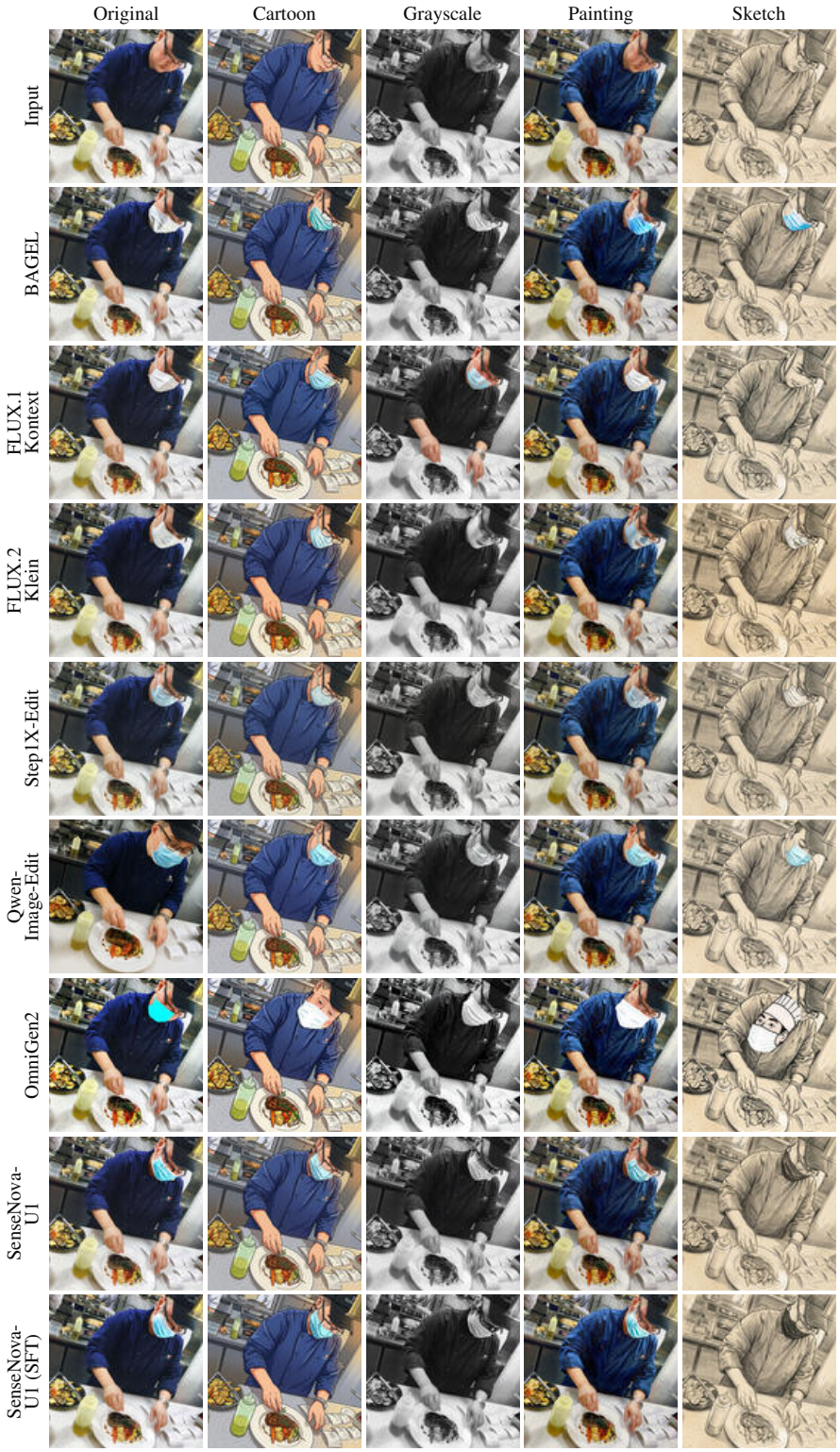


Figure 51: **Subject add examples.**

subject remove: "Remove the elderly man wearing glasses"

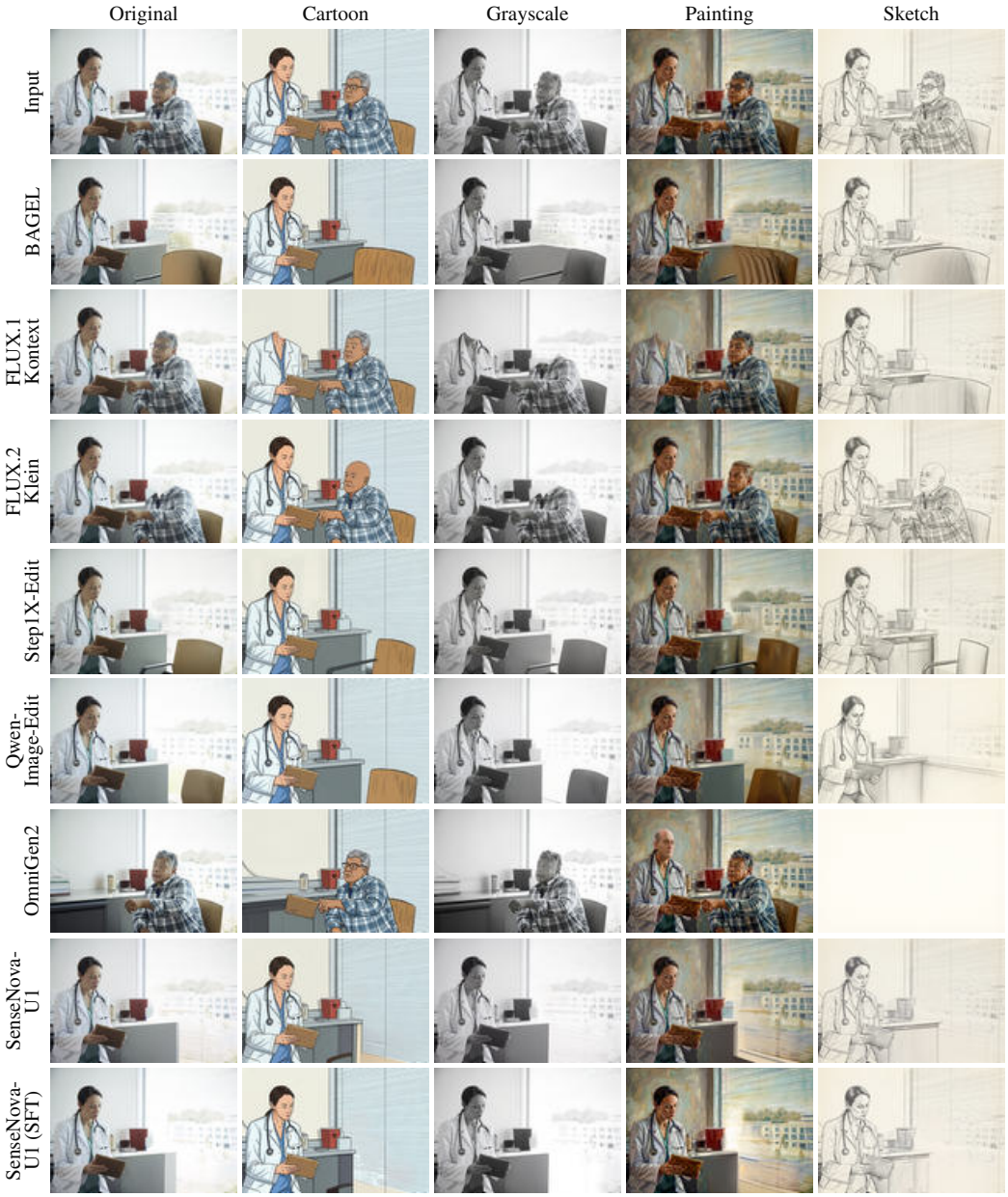


Figure 52: **Subject remove examples.**

subject replace: "Give me long hair—shoulder- length or waist-length."

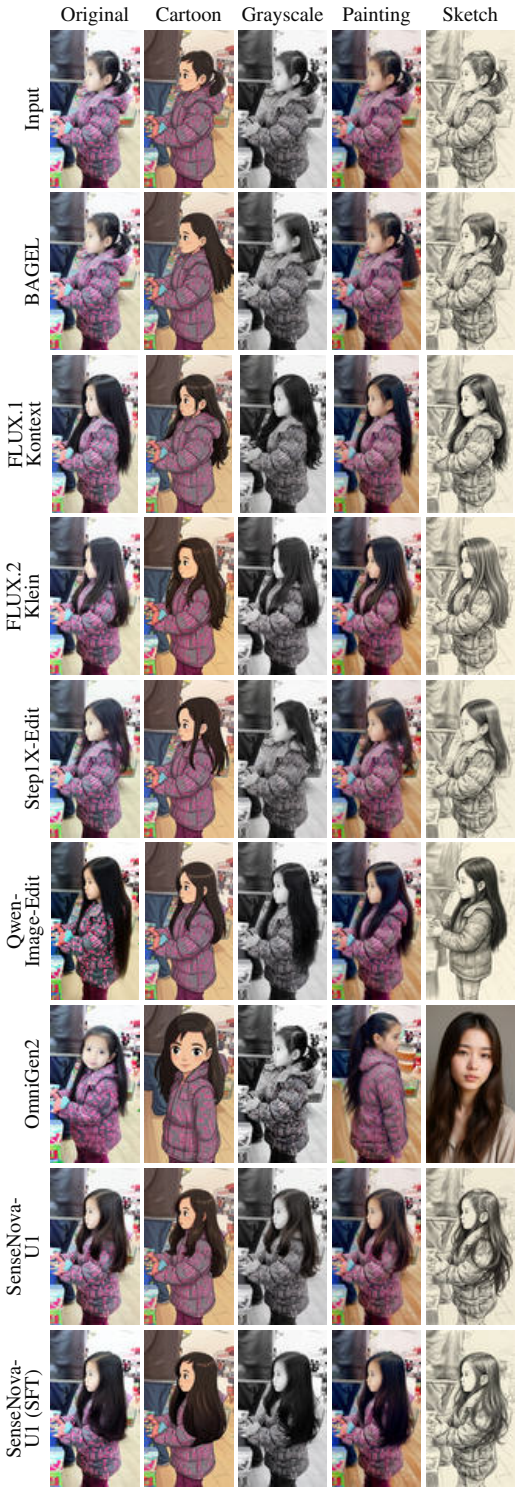


Figure 53: **Subject replace examples.**

subject replace: "Replace the eagle with a parrot."

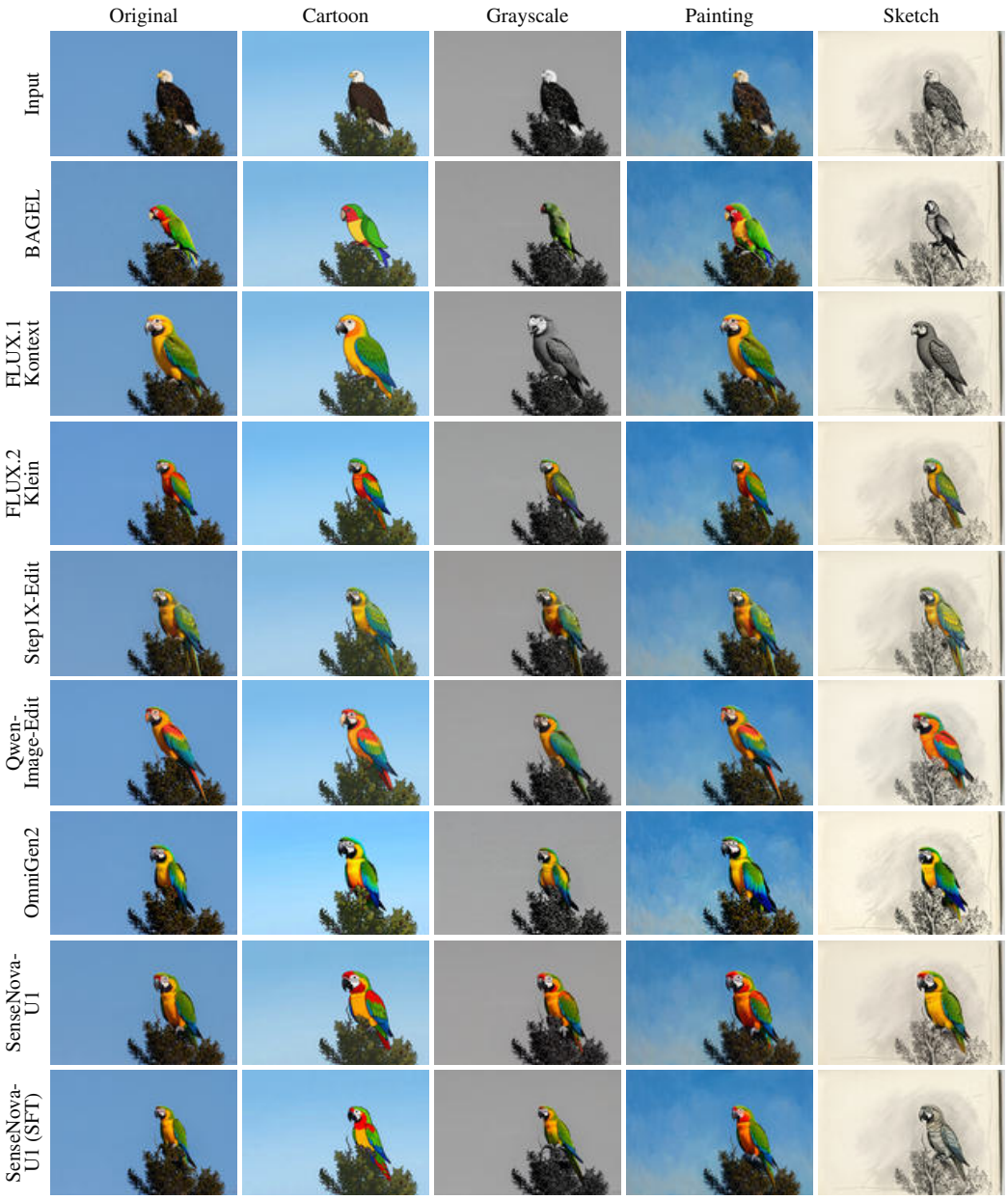


Figure 54: **Subject replace examples.**

motion change: "Create an animated version of the person in the image blinking and opening their mouth."

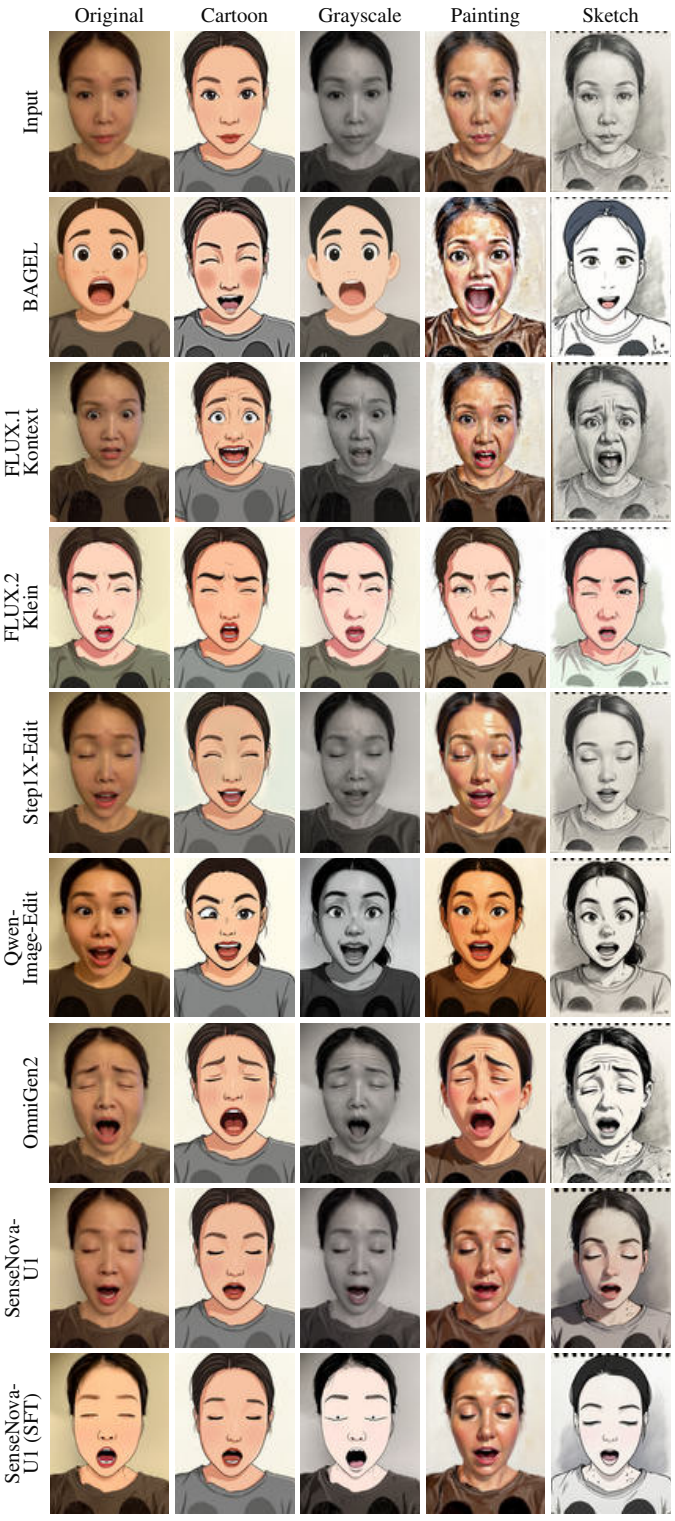


Figure 55: **Motion change examples.**