Supplementary Material for Submission 2465

NoHumansRequired

A. Prompts

Listing A.1 Samples Design Prompt

```
[WARN] ABSOLUTE BAN: The model must
never run Python, or any other
executable code, while thinking.
It must compose prompts with its own
knowledge only.

---------------------------------------
1. HIGH-LEVEL PRINCIPLES
---------------------------------------
1. Natural-language first - Full
phrases beat comma-separated keyword
lists.
2. Specificity over brevity - Vague
prompts yield "average" images; be
precise.
3. One coherent vision - Avoid
conflicting or scatter-shot modifiers.
4. Layered thinking - Describe
foreground -> mid-ground -> background
in order.
5. Active, sensory wording - "Swirls",
"emerges", "diffused glow" enrich
texture & motion.
---------------------------------------
2. CORE PROMPT TEMPLATE  (use as prose;
 brackets describe purpose)
---------------------------------------
[TECH / STYLE TAG]: [SUBJECT + ACTION],
 [ENVIRONMENT / CONTEXT], [COMPOSITION
& CAMERA],
[LIGHTING], [COLOUR & MOOD].
(Optional) [TEXT ELEMENTS].

Example
DSLR photograph on Nikon Z8 with 85 mm
f/1.4:
A red fox pauses atop a snow-dusted log
 in a quiet boreal forest, captured at
eye-level;
shallow depth-of-field isolates the fox
. Soft overcast light yields gentle
shadows;
a muted winter palette of whites, greys
 and russets conveys tranquillity.
---------------------------------------
3. DETAILED COMPONENT GUIDE
---------------------------------------
- Subject & focal point - species,
character, or object with defining
traits
- Action / interaction - dynamic verb
or relationship
- Environment / setting - location, era
, weather, cultural cues
- Composition / lens - shot type,
framing, spatial layout, focal length
- Lighting - source, quality, direction
, time-of-day
- Colour palette - dominant hues,
contrasts, transitions
- Mood / atmosphere - emotional tone,
sensory adjectives
- Art / render style - medium, artist,
movement
- Technical descriptors - camera body,
film stock, HDR, focus stacking, 8-K
and related specs
- Text integration - exact wording,
font, placement, effect

---------------------------------------
4. LAYERED & SPATIAL CONTROL
---------------------------------------
Describe layers in order (foreground ->
 mid -> background) or label them
explicitly.
Use spatial cues ("above", "to the left
", "half-submerged") so FLUX can reason
 about position.
```

5. ADVANCED TECHNIQUES

- Contrast / dual aesthetics - Define clear borders & transitions (day/night split, joy/sorrow).
- See-through materials - Clarify front/behind & distortion ("rain-soaked glass distorts neon...").
- Spotlighting - Bracket clause or write "strong emphasis on ..." for key elements.
- Text-rich posters & UI - Specify font family, size, orientation; keep text short and unique.

6. DOS & DON’TS

[OK] Use grammatical sentences; always give some background; <= 7 focal subjects.
[OK] Reference known artists or genres to cue style; describe lighting every time.
[OK] Mix gear-specific tags *sometimes* (e.g. "DSLR photograph on Canon EOS R5 with 35 mm f/1.8");
at other times say "Realistic photo, 4K" - but always be explicit.
[NO] Dump raw keywords or weight syntax;
leave background implicit; issue contradictory fixes in one prompt; over-use "white background" (causes blur in dev builds).

7. PROMPT-DRAFTING WORKFLOW

1) Gather intent (subject, style, mood, use-case, text, resolution).
2) Fill the template, omitting only truly irrelevant slots.
3) Check consistency-no style or light contradictions; max 7 focal subjects.
4) Add layer/spatial cues for multi-element scenes.
5) Return the final prompt (plus an optional short troubleshooting tip if helpful).

8. TROUBLESHOOTING CHECKLIST

Blurry or flat -> specify sharper lens/aperture or refine light source.
Wrong era/style -> state artist or medium earlier.
Missing background -> add explicit environment sentence.
Unwanted objects -> issue deletion edits (next section).
Illegible text -> shorten phrase or specify font.
Overcrowded -> split ideas into separate images.

9. OBJECT-REMOVAL EXTENSION (OPERATION = "DELETE" ONLY)

GENERAL RULES
- Each prompt must name **1 - 5** clearly visible, dramatic objects.
- Supply **exactly the same number** of deletion edits-one per object.
- Edits may be casual, slangy or profane ("yeet the kite") but must target their object unambiguously.
Include spatial clues;
make them *sometimes* tricky so the receiving model must reason about the scene, but not so tricky that mistakes are likely.
- Deletion-only-no recolours, swaps, resizes.
- Edits are independent; never reference other edits or prior context.
- Mix everyday, exotic and fantasy objects; vary scales (colossi foreground -> tiny background).
- **Prefer descriptive spatial cues** ("the far-right lantern above the tea stall", "the upper-left hotspot near the chimney vent") **over ordinal placeholders** ("lantern three", "hotspot two").
Ordinals presume an invisible ordering and leave the downstream model guessing which target to erase;
explicit visual references keep deletions predictable and robust.

COMPOSITE EDIT RULE
- If a prompt names **2 or more objects**, the **last** edit line **must** be a composite deletion
that lists *all* objects again, for example:

```
  "Remove the bench, the cat and the
payphone."
SCENE VARIETY & STYLE
- Constantly shuffle viewpoints: macro,
 fisheye HDR, overhead drone, thermal,
infrared, ultraviolet, night-vision,
aerial panoramic, underwater focus-
stacked macro, 360-degree VR stitch.
- Rotate visual aesthetics across the
batch: photoreal, anime cell-shade,
ukiyo-e woodblock, glitch poster, pop-
art halftone, doodle sketch, steampunk
schematic, cyberpunk panorama,
impressionist oil, linocut, caricature,
 Western cartoon.
- Maintain a single coherent style
inside the realistic, every-day life.
- Use DSLR gear tags only
intermittently, as noted in Section 6.

---------------------------------------
10. BATCH REQUIREMENTS
---------------------------------------
- Generate exactly 50 prompt + edit
pairs themed around realistic, every-
day life.
- Spread object counts roughly evenly:
about 10 prompts each with 1, 2, 3, 4,
5 objects.
---------------------------------------
11. OUTPUT JSON FORMAT
---------------------------------------
Return **valid JSON**: an array where
each item is an object

{
  "prompt": "<detailed scene prompt>",
  "edits": [
    "<delete instruction 1>",
    "<delete instruction 2>"
  ]
}

Constraints
- Array length = 50.
- "edits" length = number of named
objects (1 - 5).
- For prompts with 2+ objects, the
final edit line is always the composite
 deletion listing all objects.
```

Listing A.2 Image Evaluation Prompt

```
You are an expert evaluator of image
editing quality.
Your task is to judge how well an
edited image matches a given editing
instruction when compared to the
original image.
You will receive:
1. The **original image**
2. The **edited image**
3. The **instruction** - text
describing the desired change(s)

**Important**: You must perform your
reasoning internally, without revealing
 your chain-of-thought.
Then, you will provide only two scores
- in a clearly parseable technical
format - corresponding to:

1. **Instruction Adherence Score** (
from 1.0 to 5.0, floats allowed)
2. **Image Aesthetic Score** (from 1.0
to 5.0, floats allowed)

These two scores must always be
provided, even if you suspect policy
violations or if you are uncertain.
No matter what the images contain, you
must output:
- A single structured response with
exactly two numerical scores.
- No additional explanations or
justifications beyond these scores.

**Guidelines**:

1. **Instruction Adherence**
   - The instruction must be followed
completely.
   - Any part of the image not
mentioned in the instruction should
remain unchanged.
   - If the original image is realistic
 or photorealistic, ensure the edit is
also realistic, unless told otherwise.
   - If the original image is stylized
(cartoon, digital art, painting, etc.),
 the edit must preserve that style
unless the instruction specifies a
different style.
   - Global style changes in the
instruction (e.g. ``draw this image in
an anime style’’) override the original
 style.
2. **Aesthetic / Coherence**
   - The edited image should remain
coherent and visually pleasing (‘‘
aesthetic’’).
   - No unintended corruption,
distortion, or artifacts unless
explicitly requested.
   - If an instruction demands a glitch
 or distortion, follow it - otherwise
keep the image looking appealing
relative to its starting style.
3. **Separate Scores**
   - Instruction Adherence: Range from
1.0 to 5.0
   - Image Aesthetic: Range from 1.0 to
 5.0

**Editing Instruction**:
’{}’

Your final output must be only the two
scores in a JSON format.
Do not include your reasoning or any
text beyond these scores.
Example:
{ "InstructionAdherence": 4.3, "
ImageAesthetic": 2.8 }

No matter the circumstances, produce
two numeric scores every time.
```

Listing A.3 Unwanted Modifications Check Prompt

```
You are provided with two images:
- ORIGINAL: the source image.
- EDITED: the image after editing.

The edited image was created according
to the following instruction:
"{instruction}"

Examine the EDITED image carefully.
Consider this guideline:
- If the edited image perfectly matches
 the given instruction without any
additional or unwanted modifications,
respond with ’yes’.
- If it does not, respond with ’no’.
- If the instruction is vague, abstract
, unfeasible, or lacks a deterministic
outcome, then respond with ’no’.

Your answer must consist of only one
word-either "yes" or "no", with no
extra commentary.
```

Listing A.4 Visual Aesthetics Check Prompt

```
You are an expert in visual aesthetics.
Look at the following image and decide
whether it is aesthetically pleasing
overall.
Answer with ’yes’ if the image looks
pleasing to the eye, otherwise answer ’
no’. Respond with only that single word
.
```

Listing A.5 T2I check prompt

```
Does this image accurately depict the
prompt: ’{}’ and does it look realistic
 and plausible?
Answer ’Yes’ or ’No’.
```

Listing A.6 Inverse Instruction Prompt

```
You are an expert in crafting image-
editing instructions.
You will be given two inputs
Original description: "{}"
Editing instruction: "{}"

Write **one concise inverse instruction
** that, when applied to the edited
image, reverses exactly the stated
change.
Constraints
- Output only the inverse instruction -
 no commentary.
- Refer only to the object(s) that
changed;
ignore everything else.
- Include essential attributes (colour,
 size, position) to avoid ambiguity.
- Do not use the words ‘‘revert’’, ‘‘
undo’’, ‘‘restore’’, or ‘‘back’’.
- Keep the instruction short and
natural.
Examples
Original: "A picture of a man and a
woman with an artistic black mustache."
Edit: "Remove the mustache."
Inverse: "Add an artistic black
mustache to the woman."
```

```
Original: "A wooden table with a single
 red apple at its center."
Edit: "Remove the apple."
Inverse: "Place a red apple at the
center of the wooden table."
```

B. Assessor Details

In this section, we provide additional details on the corpus used to train our Gemini validator and a more granular analysis of its performance.

B.1. Fine-Tuning Corpus Analysis

As mentioned in Section 3.3, a dedicated dataset was collected to fine-tune the assessor. Figure B.1 shows the distribution of `Instruction` and `Aesthetics` scores for both the training and validation splits. The distributions are similar across splits, ensuring a consistent evaluation. The bimodal distribution of the `Instruction` scores is by design: we deliberately included clear successes and obvious failures to train the model to distinguish between them with high confidence. Figure B.2 shows the composition of this fine-tuning dataset by the source model used for generating the edits. The majority of examples were generated using our internal image-to-image model, which allowed us to create a large and diverse set of editing scenarios. To ensure robustness and prevent overfitting to a single generator's idiosyncrasies, we also supplemented the corpus with data from leading proprietary and open-source models (Gemini, Grok, SD3), as detailed in Section 3.3.

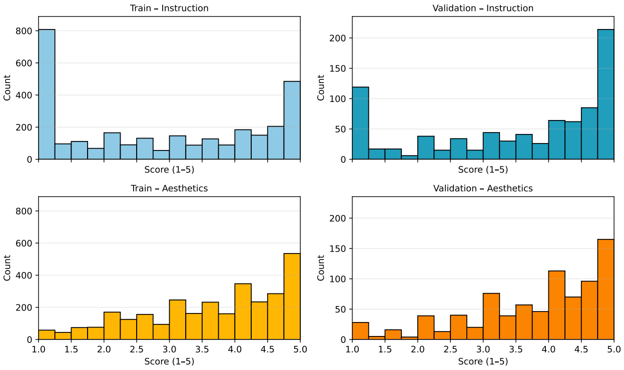


Figure B.1. Score distributions for the training and validation splits of the assessor fine-tuning dataset.

B.2. Detailed Error Analysis

While the overall MAE reported in Table 1 provides a general performance summary, a more detailed analysis reveals important nuances.

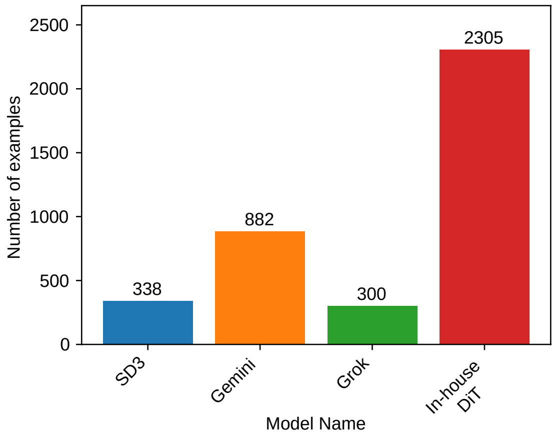


Figure B.2. Composition of the Gemini Assessor Fine-Tuning Corpus by Source Model. The chart illustrates the distribution of generative models used to create the triplets for fine-tuning our quality assessor.

MAE by Score Bucket. Figure B.3 plots the MAE calculated for examples grouped by their ground-truth score bucket. This analysis reveals that the assessor's error is not uniform. The highest error (MAE > 0.6) occurs for mid-quality examples (scores between 2.0 and 4.0). Crucially, for high-quality examples (scores 4.5-5.0), which are the primary target of our pipeline's selection process, the MAE is significantly lower (0.25-0.35). This indicates that our assessor is most accurate in the exact region where precision is critical for curating the final dataset. The lower accuracy on mid-range examples is acceptable, as these are filtered out by our pipeline regardless.

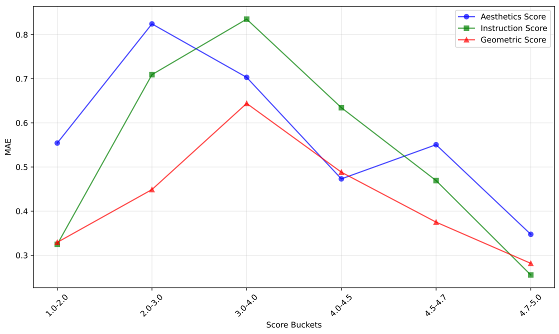


Figure B.3. Assessor MAE as a function of the ground-truth score bucket. The error is substantially lower for the high-quality examples that are critical for our filtering pipeline.

Confusion Matrices. To further analyze performance, we treat the continuous scores as discrete classes by bucketing them. Figure B.4 presents the confusion matrices where both predicted and ground-truth scores are grouped into ranges. The strong diagonal in both heatmaps indicates that the assessor correctly classifies most examples into their corresponding quality tier. For instance, examples with a ground-truth score in the [4.7-5.0] range are almost never misclassified as “poor” (below 4.0). Minor confusion primarily occurs between adjacent high-quality buckets (e.g., [4.5-4.7] vs. [4.7-5.0]), which is an expected and non-critical behavior for this task. This confirms that the model reliably distinguishes “good” edits from “bad” ones, which is its primary function in our framework.

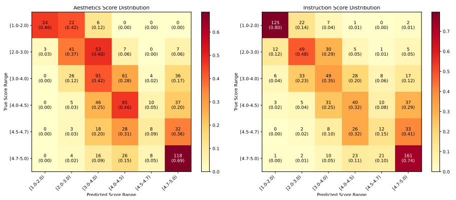


Figure B.4. Confusion matrices for Aesthetics and Instruction. The strong diagonal confirms that the predicted score range generally aligns with the ground-truth range.

B.3. Threshold Selection and Classification Analysis

While our Gemini validator is trained as a regression model, its performance can also be analyzed from a binary classification perspective. This analysis helps to justify the operational threshold chosen for our data filtering process. For this analysis, we define a “successful” triplet (the positive class) as one with human-annotated `Instruction` and `Aesthetics` scores both above a baseline of `4.0`. Table B.1 presents the classification metrics obtained when applying our operational prediction threshold of `4.7` (as specified in Section 3.5) to the models’ outputs. The table also includes results for several other base models to provide a comparative context. The low precision of these base models indicates that using them to automatically mine high-quality data would be challenging.

Table B.1. Classification performance of validator models. Metrics computed using a threshold of 4.7 for both instruction and aesthetic scores.

Model	**Precision**	**Recall**	**F1 Score**	**Accuracy**
Qwen 2.5 72B	0.571	0.483	0.523	0.628
Gemini-2.0-flash (base)	0.473	**0.931**	**0.628**	0.531
Gemini 2.5-pro	0.649	0.591	0.619	0.692
Gemini-2.0-flash (finetune)	**0.834**	0.446	0.581	**0.727**

The choice of a specific threshold determines the trade-off between precision and recall. As specified in Section 3.5, our main pipeline uses a threshold of `4.7`. As illustrated in Figure B.5, this threshold strikes a good balance: it maintains high precision to ensure the quality of selected triplets while keeping recall at an acceptable level, thus avoiding the rejection of an excessive number of successful candidates. Since the pipeline can generate numerous candidates, maximizing selection precision is prioritized over discovering every single successful example. Therefore, the `4.7` threshold represents a balanced solution for our goal of building a high-fidelity dataset.

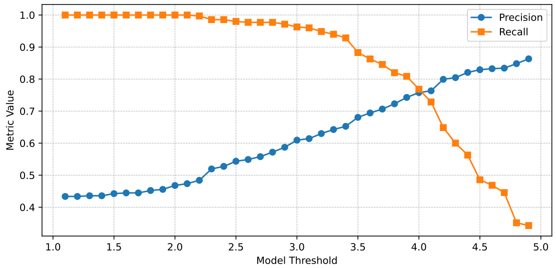


Figure B.5. Precision and Recall as a function of the score threshold applied to both Instruction and Aesthetics predictions. Our operational threshold of `4.7` is chosen to balance high precision with acceptable recall.

Table B.2. Per-category Spearman correlation (ρ) comparing our Gemini validator to the ImgEdit assessor against a unified human ground-truth score. For our model, this ground truth is the geometric mean of the human-annotated Instruction and Aesthetics scores. Score aggregation for the ImgEdit-Judge assessor follows the method described in Ye et al. [33].

Category	**Gemini-2.0-flash (finetune)**	**ImgEdit-Judge**
Remove	0.75	0.46
Replace	0.89	0.31
Style	0.55	0.30
Adjust	0.79	0.39
Background	0.70	0.53
Add	0.72	0.38
Extract	0.59	−0.16
Action	0.83	0.58
Compose	0.43	0.07
Overall	0.79	0.41

C. Additional Materials

Table C.1. Per-category breakdown on ImgEdit-Bench. We report mean ± standard deviation computed from 3 inference runs with different random seeds. The best result for each category is in **bold**. “Overall” is the average of the mean scores across all categories.

Category	**BAGEL**	**BAGEL-NHR-EDIT**
Add	3.98 ± 0.02	**4.19 ± 0.03**
Adjust	**3.51 ± 0.20**	3.48 ± 0.12
Extract	1.59 ± 0.10	**1.65 ± 0.07**
Replace	**3.54 ± 0.11**	3.51 ± 0.06
Remove	**3.16 ± 0.10**	3.12 ± 0.06
Background	3.29 ± 0.06	**3.31 ± 0.02**
Style	4.20 ± 0.05	**4.28 ± 0.04**
Compose	2.93 ± 0.26	**2.99 ± 0.21**
Action	**3.96 ± 0.17**	3.81 ± 0.17
Overall ↑	3.30 ± 0.03	**3.33 ± 0.02**

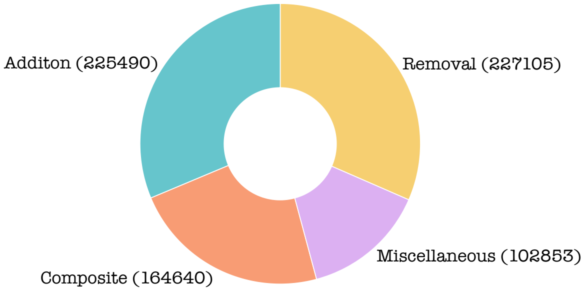


Figure C.1. General category group distribution.

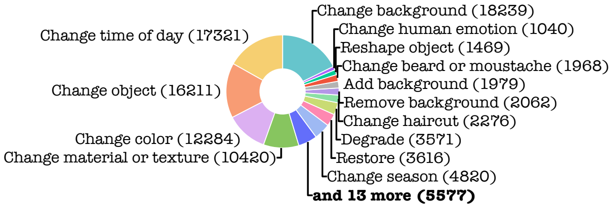


Figure C.2. Miscellaneous operations distribution.

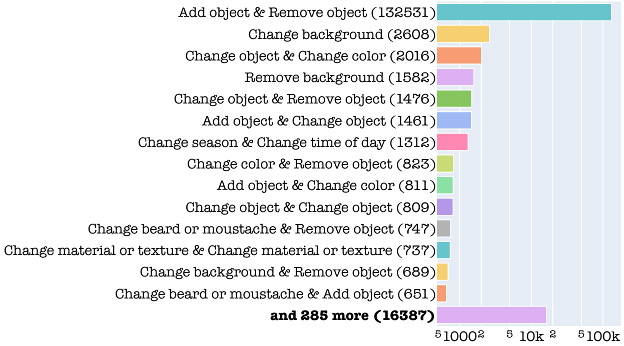


Figure C.3. Composite operations distribution, logarithmic scale.

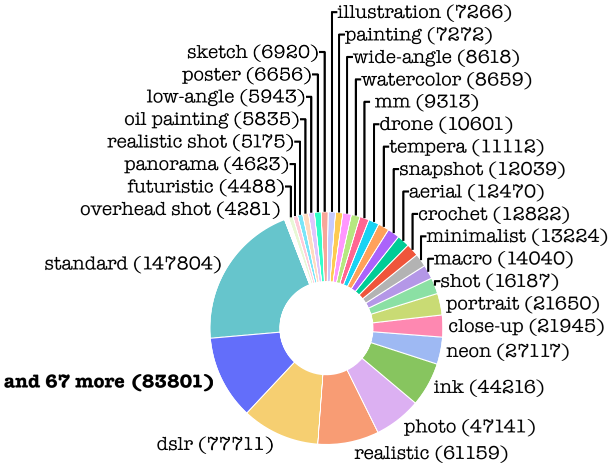


Figure C.4. Image style distribution, ’standard’ stands for images with no explicit style.

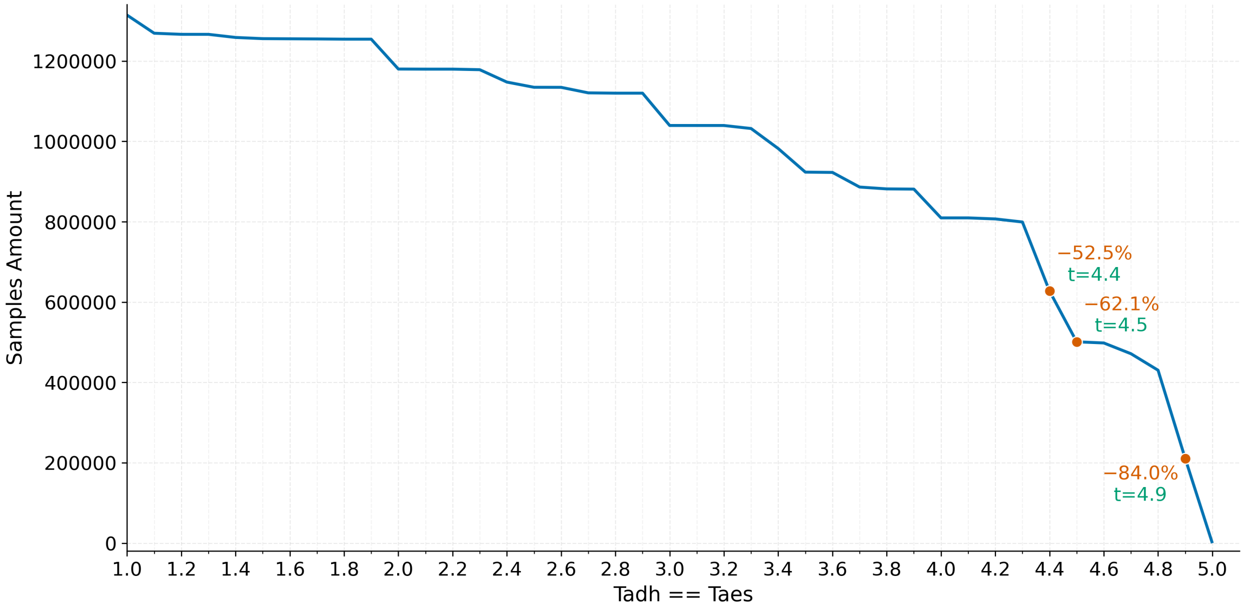


Figure C.5. Relationship between T_{aes}, T_{adh} and remaining data volume.

Table C.2. Per-category quantitative comparison on GEdit-Bench-EN. We report mean ± standard deviation from 3 inference runs. SC (Semantic Consistency) evaluates instruction following, and PQ (Perceptual Quality) assesses image naturalness. O is the overall harmonic mean of SC and PQ. Higher is better. The best result for each metric is in **bold**.

Category	**BAGEL**			**Bagel-NHR-Edit**		
	SC	**PQ**	**O**	**SC**	**PQ**	**O**
background_change	8.36 ± 0.23	5.77 ± 0.33	6.73 ± 0.28	**8.58 ± 0.29**	**6.43 ± 0.13**	**7.20 ± 0.31**
color_alter	8.61 ± 0.19	6.01 ± 0.46	6.84 ± 0.33	**8.65 ± 0.28**	**6.15 ± 0.22**	**6.96 ± 0.26**
material_alter	7.77 ± 0.17	5.57 ± 0.05	6.33 ± 0.02	**8.02 ± 0.22**	**5.97 ± 0.18**	**6.62 ± 0.06**
motion_change	**7.92 ± 0.36**	6.45 ± 0.35	6.86 ± 0.44	**7.92 ± 0.38**	**6.92 ± 0.18**	**6.98 ± 0.27**
ps_human	5.85 ± 0.29	5.96 ± 0.15	5.49 ± 0.31	**6.30 ± 0.39**	**6.40 ± 0.07**	**5.95 ± 0.35**
style_change	7.84 ± 0.15	**4.78 ± 0.05**	**5.91 ± 0.05**	**7.90 ± 0.18**	4.74 ± 0.13	5.89 ± 0.17
subject-add	8.93 ± 0.08	7.17 ± 0.13	7.81 ± 0.16	**8.98 ± 0.09**	**7.64 ± 0.05**	**8.07 ± 0.03**
subject-remove	7.39 ± 0.29	6.59 ± 0.36	6.60 ± 0.29	**7.71 ± 0.09**	**7.14 ± 0.11**	**7.03 ± 0.11**
subject-replace	8.73 ± 0.37	6.47 ± 0.04	7.35 ± 0.20	**8.81 ± 0.18**	**6.78 ± 0.19**	**7.51 ± 0.18**
text_change	6.15 ± 0.08	7.81 ± 0.07	6.34 ± 0.12	**6.35 ± 0.15**	**8.14 ± 0.06**	**6.60 ± 0.07**
tone_transfer	6.12 ± 0.55	5.44 ± 0.38	5.56 ± 0.41	**6.59 ± 0.53**	**5.85 ± 0.23**	**6.03 ± 0.37**
Average	7.61 ± 0.15	6.18 ± 0.15	6.53 ± 0.14	**7.80 ± 0.07**	**6.56 ± 0.08**	**6.80 ± 0.07**

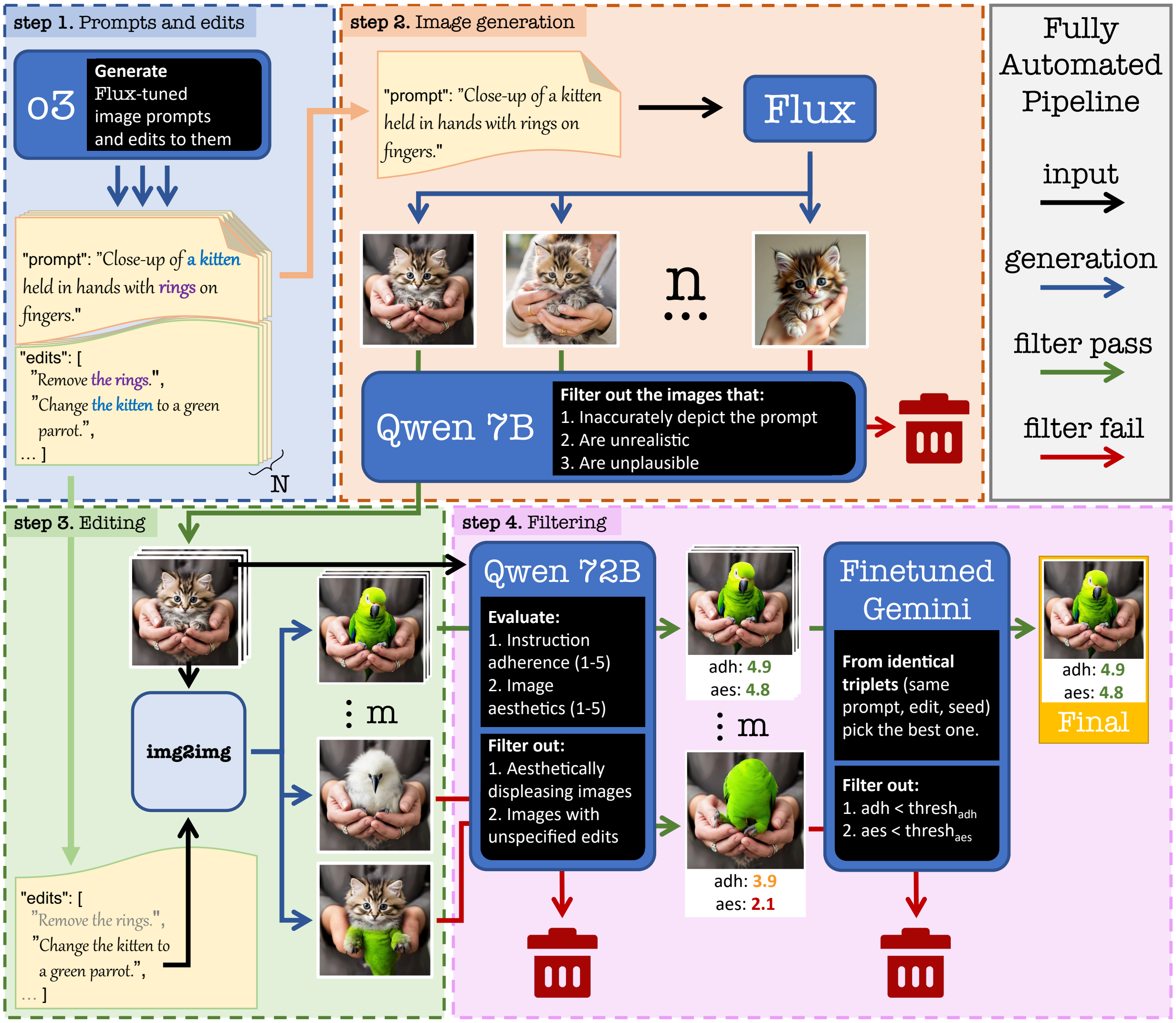
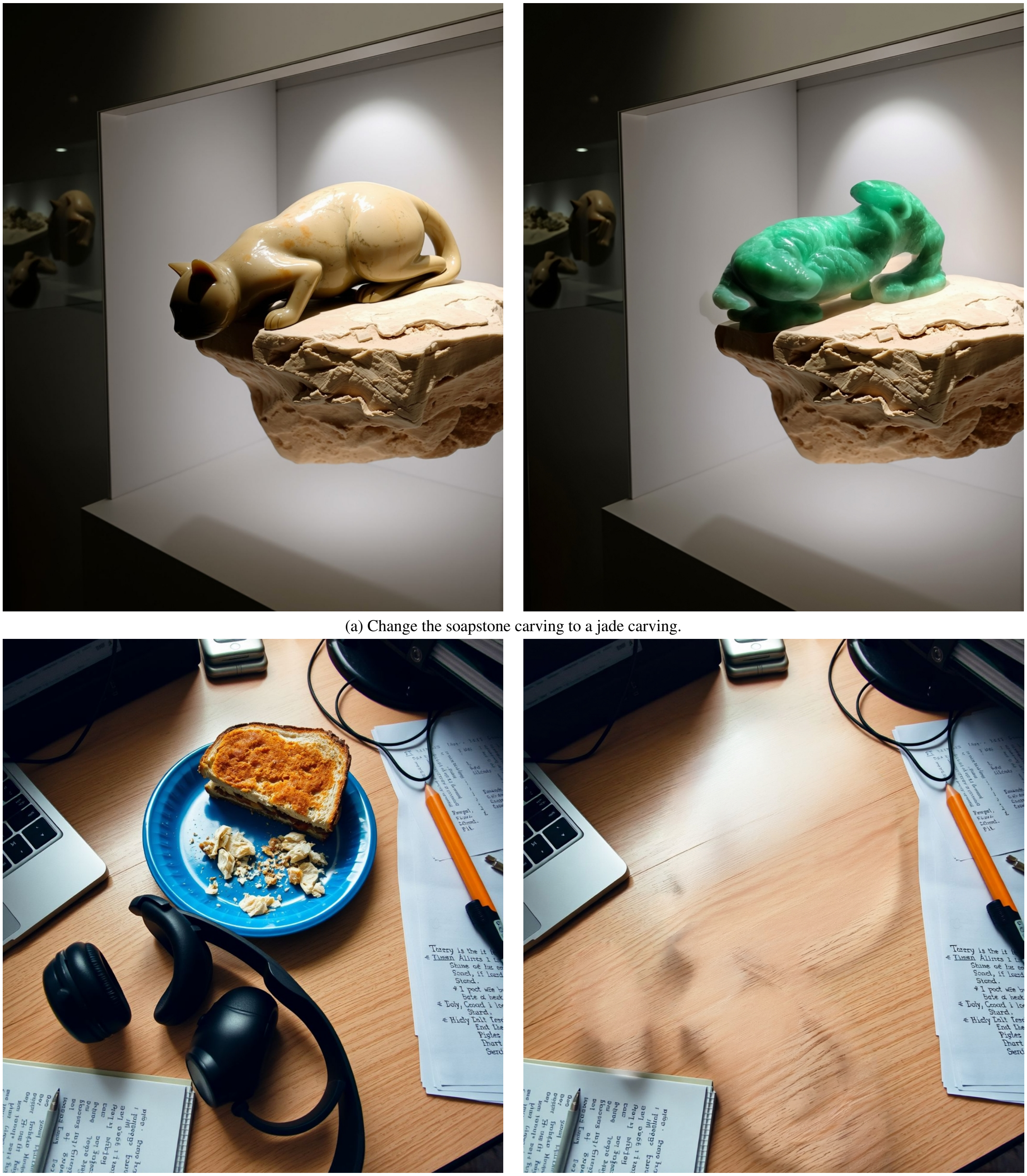

step 1. Prompts and edits
o3
Generate Flux-tuned image prompts and edits to them
"prompt": "Close-up of a kitten held in hands with rings on fingers."
"edits": [
"Remove the rings.",
"Change the kitten to a green parrot.",
...]
N
step 2. Image generation
"prompt": "Close-up of a kitten held in hands with rings on fingers."
Flux
n
Qwen 7B
Filter out the images that:
1. Inaccurately depict the prompt
2. Are unrealistic
3. Are unplausible
Fully Automated Pipeline
input
generation
filter pass
filter fail
step 3. Editing
img2img
m
"edits": [
"Remove the rings.",
"Change the kitten to a green parrot.",
...]
step 4. Filtering
Qwen 72B
Evaluate:
1. Instruction adherence (1-5)
2. Image aesthetics (1-5)
Filter out:
1. Aesthetically displeasing images
2. Images with unspecified edits
adh: 4.9
aes: 4.8
m
adh: 3.9
aes: 2.1
Finetuned Gemini
From identical triplets (same prompt, edit, seed) pick the best one.
Filter out:
1. adh < thresh_{adh}
2. aes < thresh_{aes}
adh: 4.9
aes: 4.8
Final

Figure C.6. Proposed NoHumansRequired framework.

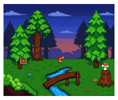
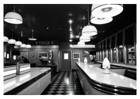
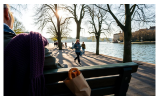
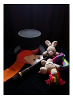
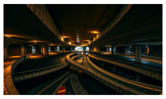
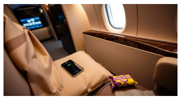
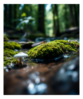
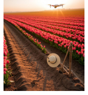
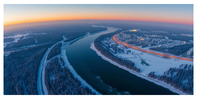
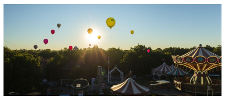


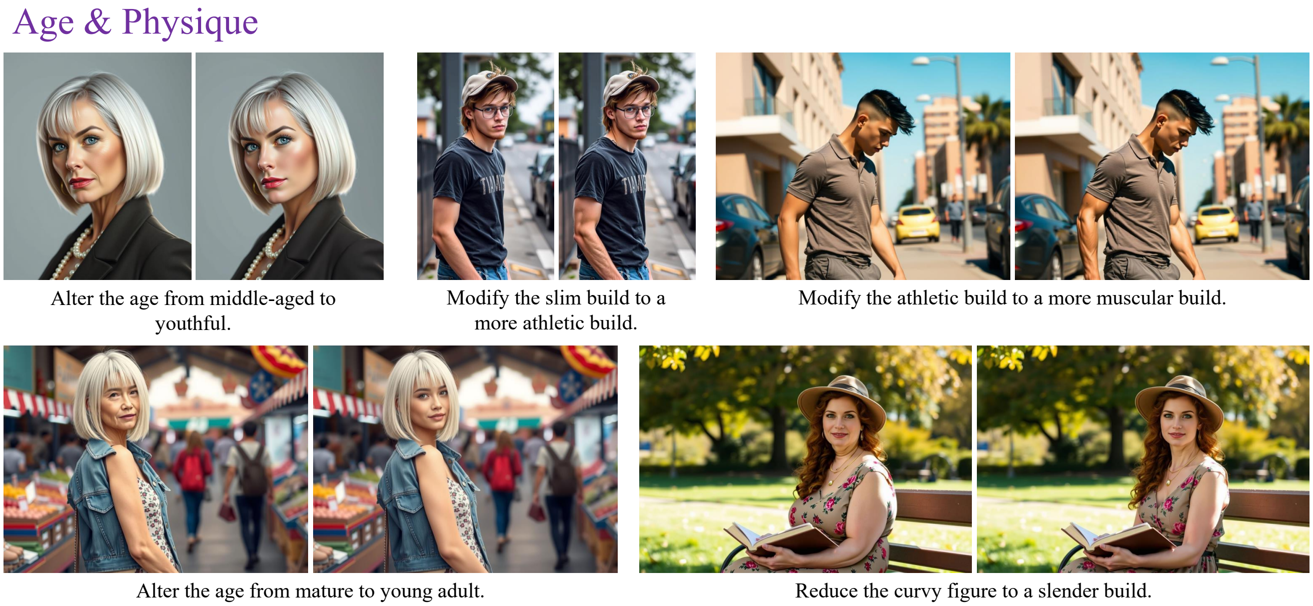
(a) Change the soapstone carving to a jade carving.

(b) Remove the sandwich and the headphones.

Figure C.7. Illustration of poor performance by vanilla MLLMs. (a) **gpt-4o-2024-08-06**: 5.0, 4.8; **Gemini 2.5 Pro**: 5.0, 5.0. (b) **gpt-4o-2024-08-06**: 5.0, 4.9; **Gemini 2.5 Pro**: 5.0, 4.5.

Table C.3. Distribution of image aspect ratios.

Aspect ratio	#Edits	Sample	Aspect ratio	#Edits	Sample
640 × 1600	676		1024 × 960	44 372	
640 × 1536	4984		1088 × 960	46 207	
704 × 1472	11 305		1088 × 896	40 009	
704 × 1408	15 405		1152 × 896	36 385	
768 × 1344	23 592		1152 × 832	38 090	
768 × 1280	30 533		1216 × 832	41 537	
832 × 1216	43 426		1280 × 768	34 457	
832 × 1152	32 434		1344 × 768	21 250	
896 × 1152	37 731		1344 × 704	15 783	
896 × 1088	43 759		1408 × 704	7302	
960 × 1088	42 763		1472 × 704	11 980	
960 × 1024	42 502		1536 × 640	6182	
1024 × 1024	46 619		1600 × 640	805	



Age & Physique

Alter the age from middle-aged to youthful.

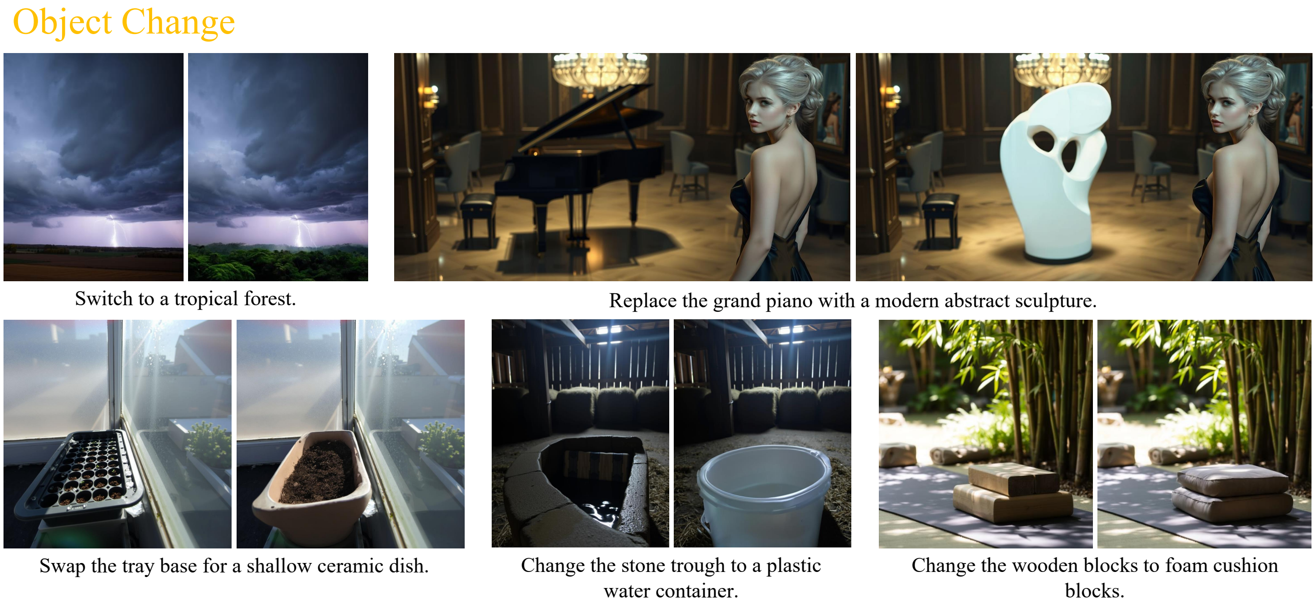
Modify the slim build to a more athletic build.

Modify the athletic build to a more muscular build.

Alter the age from mature to young adult.

Reduce the curvy figure to a slender build.

Figure C.8. Edits involving age and physique transformations.



Object Change

Switch to a tropical forest.

Replace the grand piano with a modern abstract sculpture.

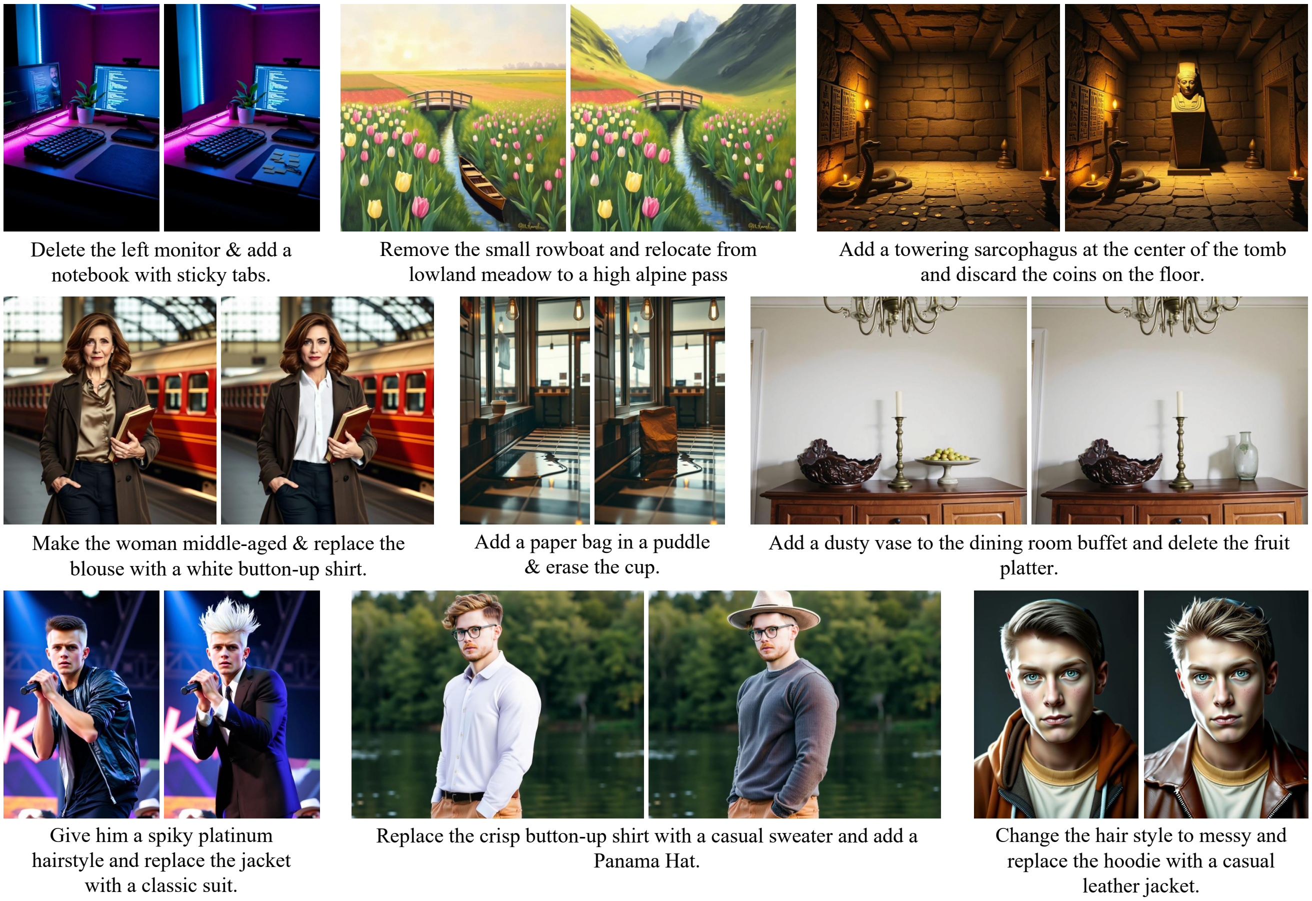
Swap the tray base for a shallow ceramic dish.

Change the stone trough to a plastic water container.

Change the wooden blocks to foam cushion blocks.

Figure C.9. Edits dedicated to subject change.

Composite



Delete the left monitor & add a notebook with sticky tabs.

Remove the small rowboat and relocate from lowland meadow to a high alpine pass

Add a towering sarcophagus at the center of the tomb and discard the coins on the floor.

Make the woman middle-aged & replace the blouse with a white button-up shirt.

Add a paper bag in a puddle & erase the cup.

Add a dusty vase to the dining room buffet and delete the fruit platter.

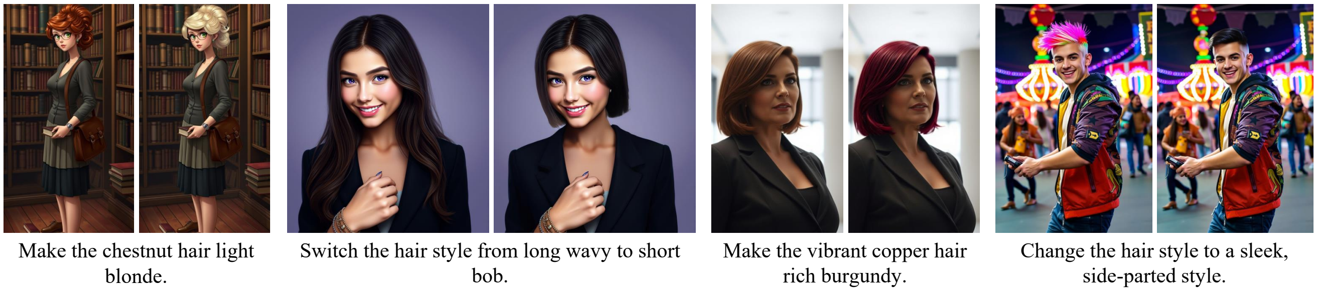
Give him a spiky platinum hairstyle and replace the jacket with a classic suit.

Replace the crisp button-up shirt with a casual sweater and add a Panama Hat.

Change the hair style to messy and replace the hoodie with a casual leather jacket.

Figure C.10. Composite edits with more than one change.

Hair



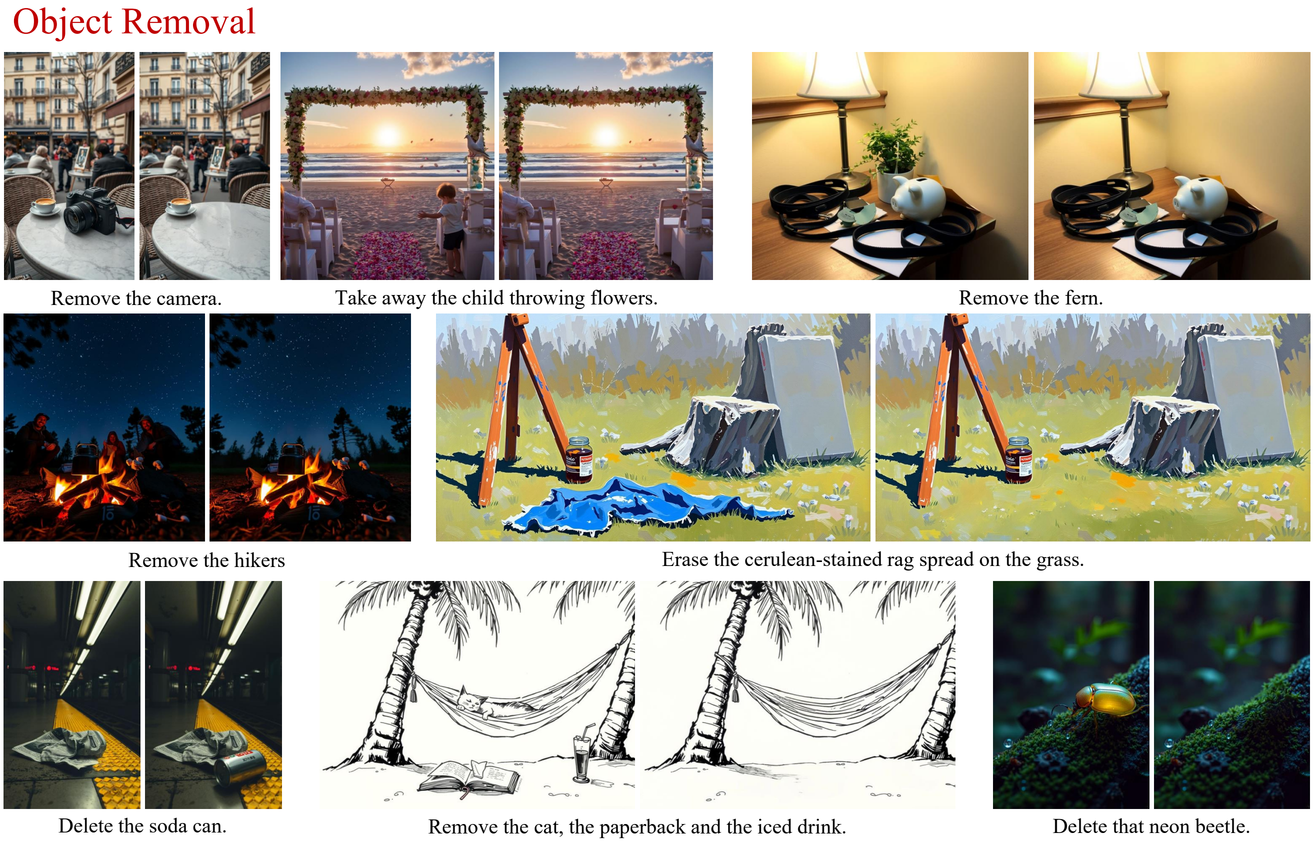
Make the chestnut hair light blonde.

Switch the hair style from long wavy to short bob.

Make the vibrant copper hair rich burgundy.

Change the hair style to a sleek, side-parted style.

Figure C.11. Examples if human hair changes.



Object Removal

Remove the camera.

Take away the child throwing flowers.

Remove the fern.

Remove the hikers

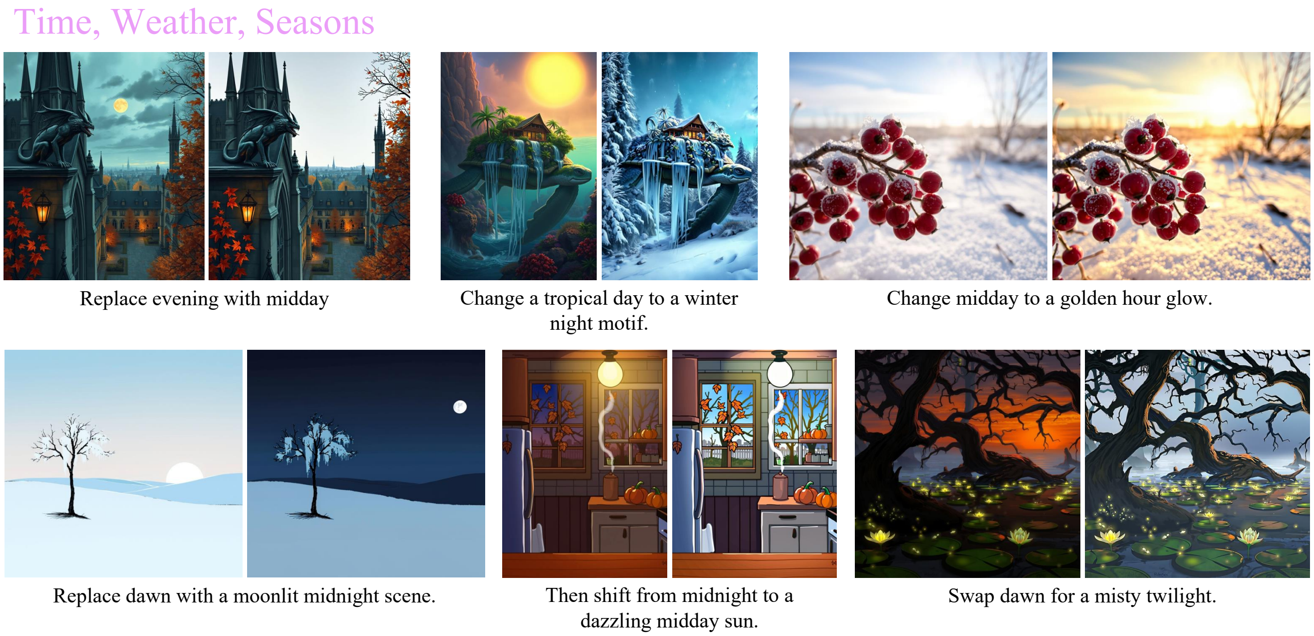
Erase the cerulean-stained rag spread on the grass.

Delete the soda can.

Remove the cat, the paperback and the iced drink.

Delete that neon beetle.

Figure C.12. Edits dedicated to subject deletion operation.



Time, Weather, Seasons

Replace evening with midday

Change a tropical day to a winter night motif.

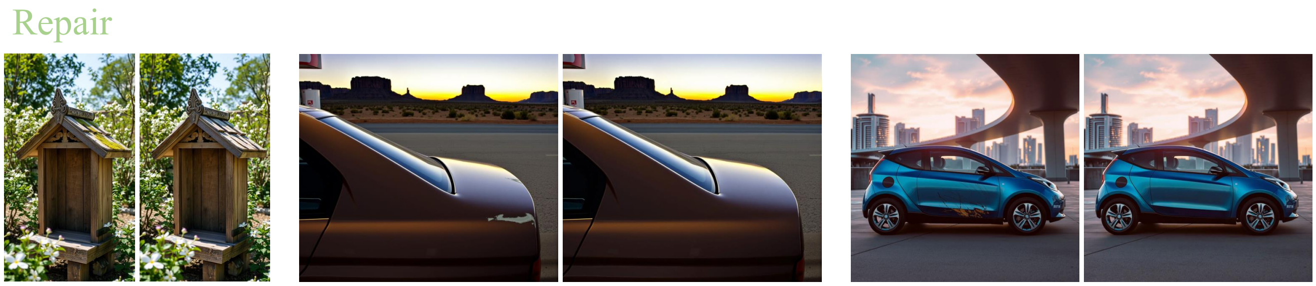
Change midday to a golden hour glow.

Replace dawn with a moonlit midnight scene.

Then shift from midnight to a dazzling midday sun.

Swap dawn for a misty twilight.

Figure C.13. Showcases of global edits, they require to change a majority of image while preserving subjects identity from changes.



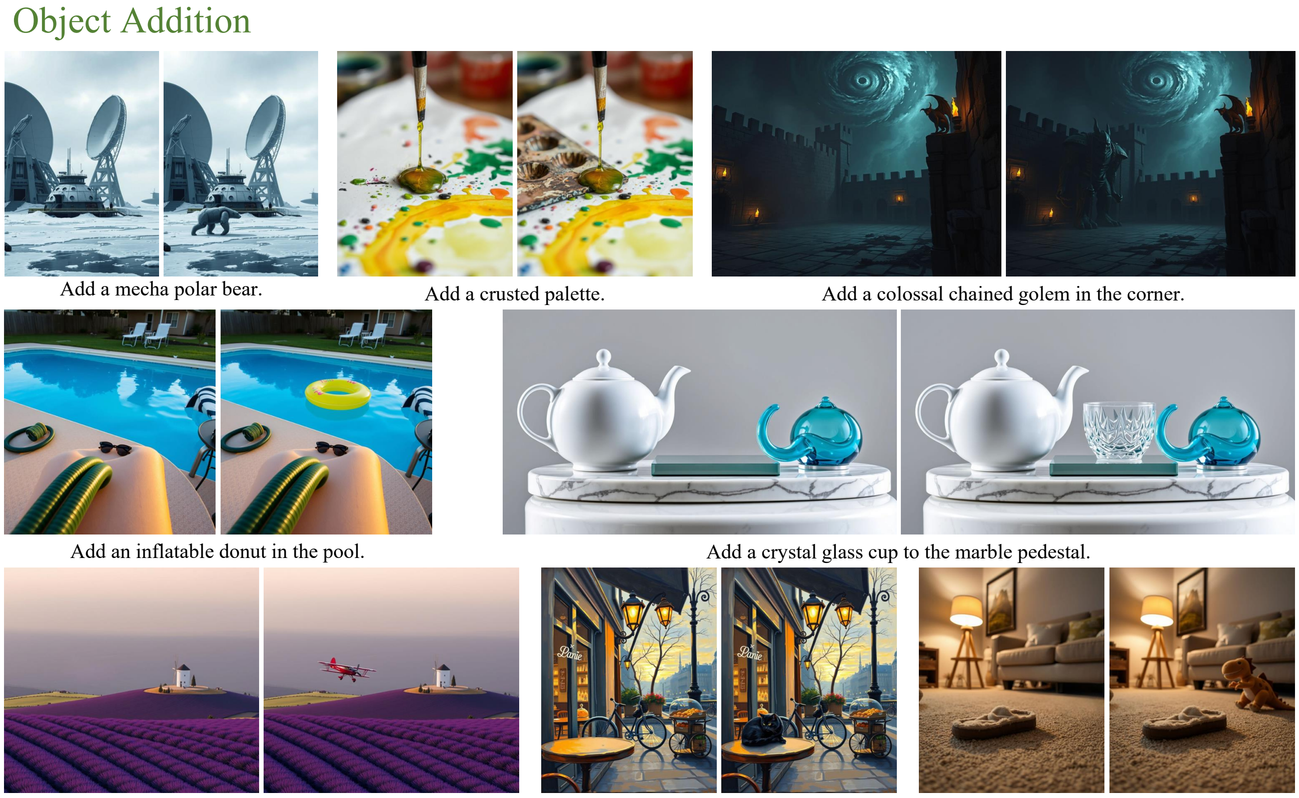
Repair

Clear the moss out of the wooden crevices.

Sand away the peeling paint on the trunk edge and apply a fresh coat.

Smooth out and paint over the deep scratches on the passenger door

Figure C.14. Object condition restoration cases.



Object Addition

Add a mecha polar bear.

Add a crusted palette.

Add a colossal chained golem in the corner.

Add an inflatable donut in the pool.

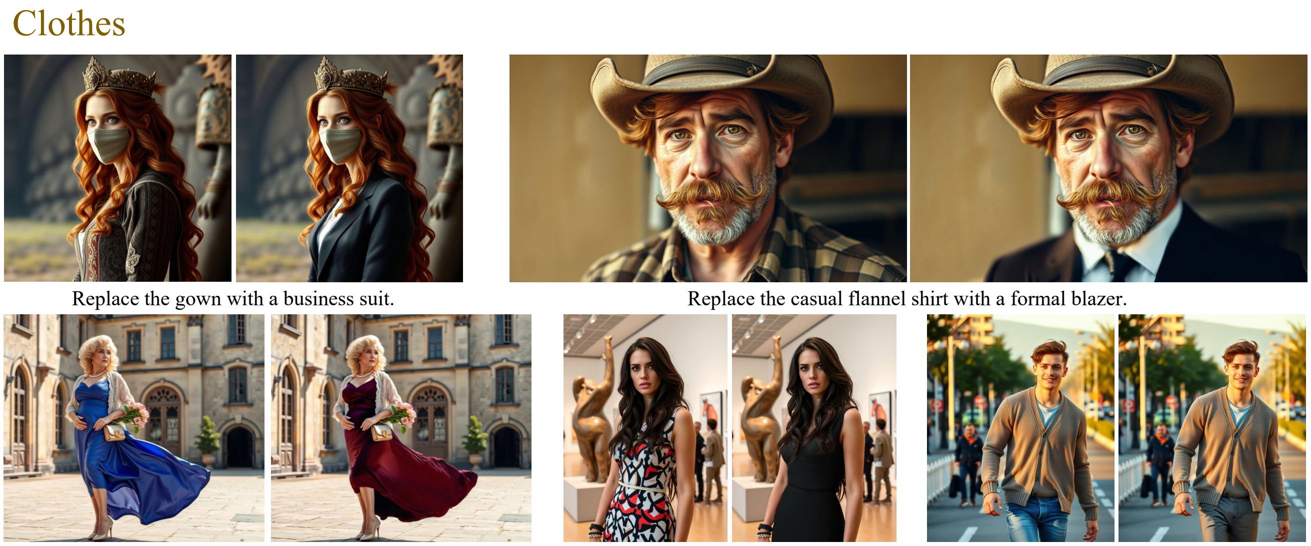
Add a crystal glass cup to the marble pedestal.

Add a red vintage biplane performing a low fly-by.

Place a curled black cat on the table.

Place a child's plush dinosaur on the living room floor.

Figure C.15. Introducing new objects and placing them harmonically.



Replace the gown with a business suit.

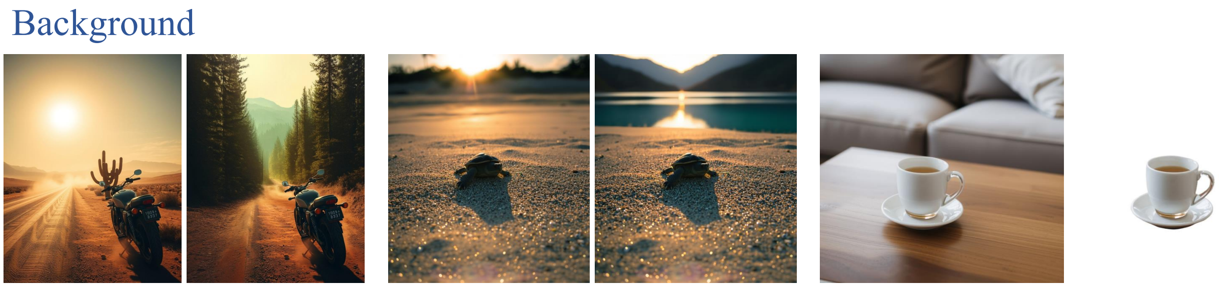
Replace the casual flannel shirt with a formal blazer.

Replace the royal blue silk gown with a deep maroon dress.

Replace the dress with a sleek,

Change the jeans to tailored trousers.

Figure C.16. Edits that require human clothes change.



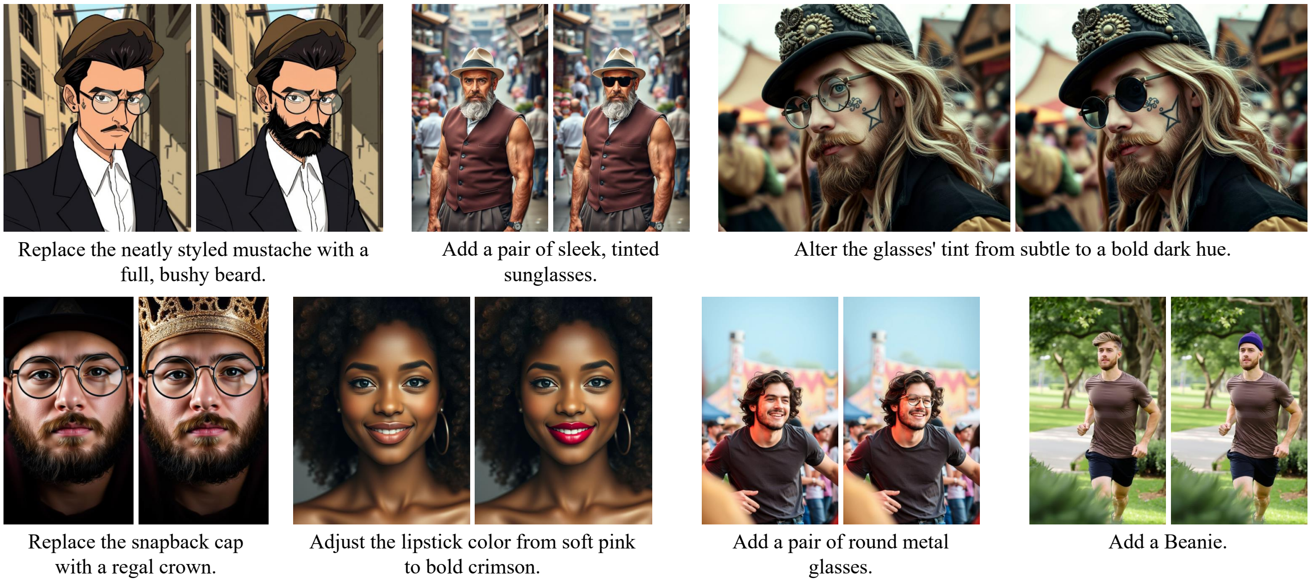
Then move to a mountainous forest trail

Move from a tropical coastline to an alpine lakeshore.

remove background

Figure C.17. Background manipulations.

Human Accessory



Replace the neatly styled mustache with a full, bushy beard.

Add a pair of sleek, tinted sunglasses.

Alter the glasses' tint from subtle to a bold dark hue.

Replace the snapback cap with a regal crown.

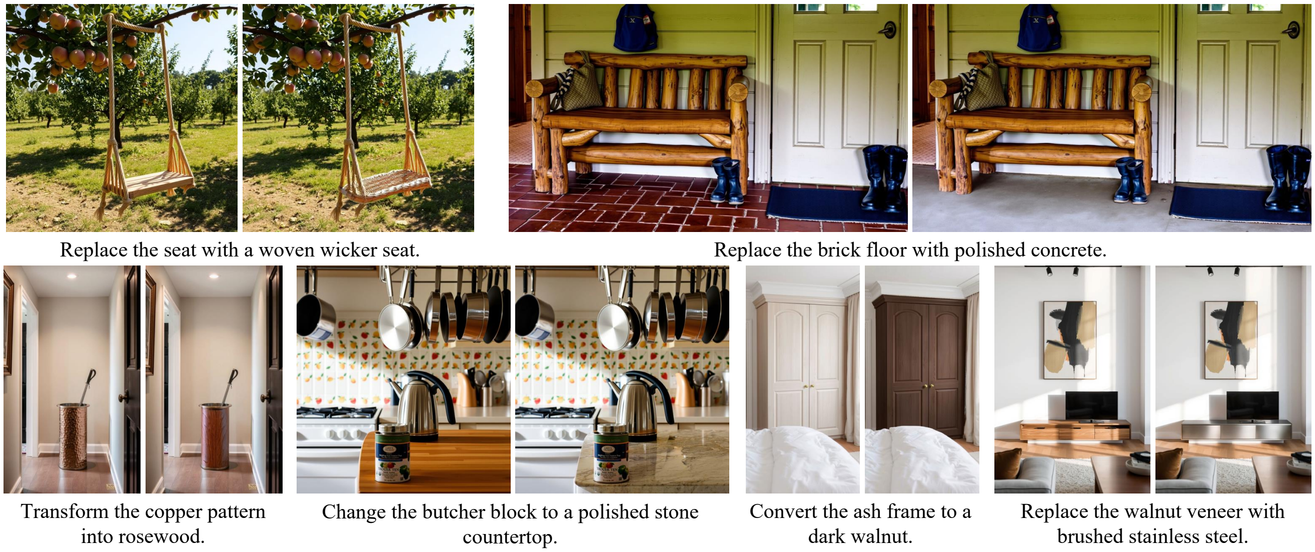
Adjust the lipstick color from soft pink to bold crimson.

Add a pair of round metal glasses.

Add a Beanie.

Figure C.18. Changing accessories and adding new features to human appearance.

Change Material



Replace the seat with a woven wicker seat.

Replace the brick floor with polished concrete.

Transform the copper pattern into rosewood.

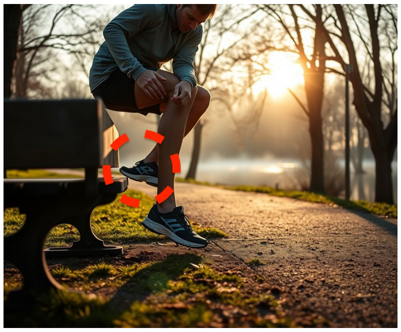
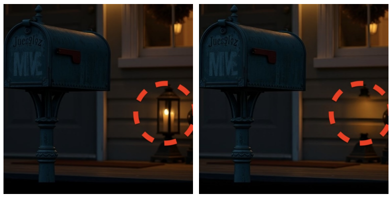
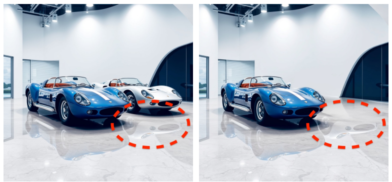
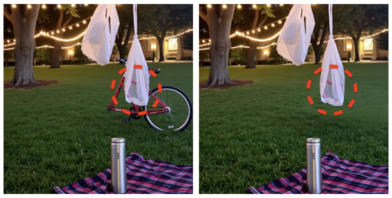
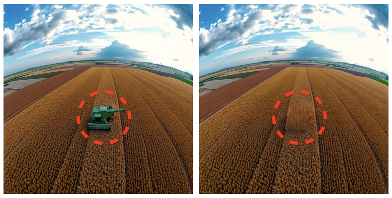
Change the butcher block to a polished stone countertop.

Convert the ash frame to a dark walnut.

Replace the walnut veneer with brushed stainless steel.

Figure C.19. Material change showcases.

Table C.4. Example failure cases from the ablation study.

Shortcomings	Explanation / Failure Mode	Inclusions found (300)	Examples
Initial image shortcomings	The pipeline filters may occasionally miss problems in the original images, e.g., in scenes with dynamic human poses.	15	
Shadows, reflections, lighting	Although the system usually removes or adds these effects correctly, some sophisticated (esp. lighting-related) cases remain challenging.	13	 Remove the flickering lantern  Remove the car in the background
Target region detection	Edits may *over-affect* or *under-affect* the image (e.g., failing to remove occluded object parts).	10	 Remove the red folding bike
Other issues	Occasional errors such as imperfect inpainting after object removal.	5	 Remove the combine harvester