A. Prompts

VLM Filtering Prompt

```
You are evaluating a low-resolution synthetic render for a captioning dataset.

Decide if the image is GOOD or BAD.

Output format (exactly one line):

- "GOOD: <brief factual reason>"
- "BAD: <brief factual reason>"

Goal: keep only images that are EASY to caption accurately (specific nouns
+ attributes).

Reject images that would force a vague caption (e.g., "a close-up of something").

PASS CONDITIONS (either is enough)
A) Object-centric: a recognizable object/character is shown with enough context
to name it.
B) Scene-centric: a recognizable scene/environment is shown (forest, room,
street, landscape).

HARD REJECT (always BAD)
1) EXTREME CROP / CLOSE-UP
- BAD if the view is an extreme close-up or partial fragment such that
  the subject cannot be confidently named.
- BAD if the frame is dominated by a single surface/part (e.g., cheek, wall,
  texture) without context.
- BAD if >30% of the subject is cut off OR the crop removes key identifying
  parts (e.g., head missing, face half missing, object mostly out of frame).
2) IDENTIFIABILITY FAILURE
- BAD if you cannot identify WHAT it is (object type OR scene type) in one
  short noun phrase.
3) RENDER / SYNTHETIC ERRORS
- BAD if obvious rendering artifacts exist: clipping/interpenetration,
  broken geometry, missing textures/materials,
  NaN/black patches, fireflies/bright speckles, extreme distortion.
4) VISIBILITY FAILURE
- BAD if too dark/bright/blurred/noisy to recognize major shapes and boundaries.
- BAD if mostly blank/black/solid color.

FRAMING RULES
- Object-centric GOOD only if the full object OR a clearly intentional,
  informative partial view is shown.
  (Example acceptable partial: "close-up of a clock face" where it is clearly
    a clock.)
- Scene-centric GOOD only if the scene layout is readable and not dominated by
  an uninformative foreground occluder.
OCCLUSION
- GOOD if occlusion is natural and still captionable.
- BAD if key content is blocked or inexplicably obscured.
INSTRUCTIONS
- Low resolution alone is NOT a reason to reject.
- Be strict: if the best caption would be vague, mark BAD.
- Keep the reason short and factual (5{10 words).
- Output exactly one line.
```

VLM Filtering Prompt

```
Describe the image with a single factual caption.

Output format:
- A single sentence caption (no extra text).

Rules:
- Describe only what is clearly visible in the image.
- Do NOT guess or hallucinate unseen objects.
- Do NOT include opinions, emotions, or storytelling.
- Use concrete nouns and simple attributes (colors, materials, positions).
- Mention the main objects and the environment if visible.
- If it is a scene (e.g., landscape, forest, room), describe the scene
  structure.
- If it is an object, mention the object and its context/background.

Style guidelines:
- 8{20 words.
- Neutral, factual tone.
- No phrases like "this image shows" or "there is".
- Avoid unnecessary adjectives.

Examples:

GOOD:
"Three trees standing on a grassy field under a cloudy sky."
"A wooden chair placed next to a small table in a bright room."
"A red car parked on a paved road beside a row of buildings."

BAD:
"This image shows a beautiful scene with some trees."
"A very nice and detailed picture of a forest landscape."
"Possibly a statue or object that looks like metal."

Return only the caption.
```