SUPPLEMENTARY MATERIALS

We provide additional experimental details in Section A. We give our user study details in Section B. We discuss the applications and implications of Stencil in Section C and provide additional qualitative results in Section D.

A. Additional Implementation Details

A.1. Generating Image-Text pairs

We use LangChain to transform GPT-4o's unstructured textual outputs into structured responses.

For each user-provided reference image, we use GPT-4o to generate a corresponding caption, forming image-text pairs which we can then use to fine-tune the U-Net backbone. We use the below system message to have GPT-4o perform the required task for us.

```
You are an professional at captioning images.
You are given some images of a subject.

You are tasked to perform the following:
1. Provide a short description of the subject
    ↪ , subject_name.
2. Create a detailed caption for each image
    ↪ containing the subject_name,
    ↪ image_caption.

You are to respond in the JSON format defined
    ↪  below.

Format Instructions:
--------------
{format_instructions}
--------------
```

In general, we find that fine-tuning reference images on descriptive captions yield more diverse results and is significantly less prone to overfitting compared to using concise prompts. We attribute this to language drift. When a prompt lacks sufficient detail, the model may inadvertently bind the subject tokens to both the subject's and the background's representation. Using a more descriptive prompt helps disentangle these features, thereby improving the model's ability to generalize. We demonstrate this point in Fig. 1, where we compare the output of a U-Net fine-tuned on a set of concise captions vs a set of detailed captions.

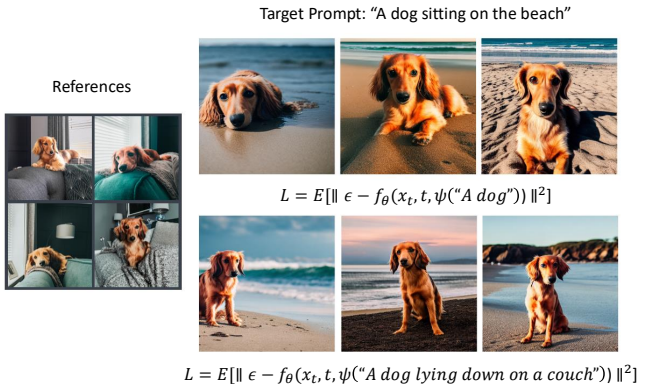

Fig. 1: **Impact of Concise vs. Detailed Prompts During Fine-Tuning.** We compare outputs from a U-Net fine-tuned on a concise caption (Top Row) versus a detailed caption (Bottom Row) of the reference images. When the caption lacks sufficient detail, the model tends to overfit to the reference image, producing less diverse generations.

A.2. Fine-tuning the Decoder

We demonstrate in Fig. 3 that as spatial features propagate through the U-Net, higher-frequency information is captured. The shallower layers of the U-Net learn the structure, whereas the deeper layers learn the finer appearances of the image. Since subject-driven generation concerns the learning of higher-frequency details (e.g., appearance, color, texture, shape, etc.), we only fine-tune the U-Net decoder blocks in our implementation while freezing the rest of the network.

A.3. Cross-Attention Guided Loss Threshold p_t

We evaluate different values of the threshold p_t to determine the optimal settings that maximizes the separability of subject and background pixels. This threshold represents the minimum attention weight a pixel must have toward the subject token to be considered relevant, and any pixels with an attention weight below this threshold are excluded from the loss computation. A threshold that is too low may include irrelevant background features in the loss computation, whereas a high threshold risks omitting important subject regions. Based on the results shown in Fig. 2, we select $p_t = 0.2$.

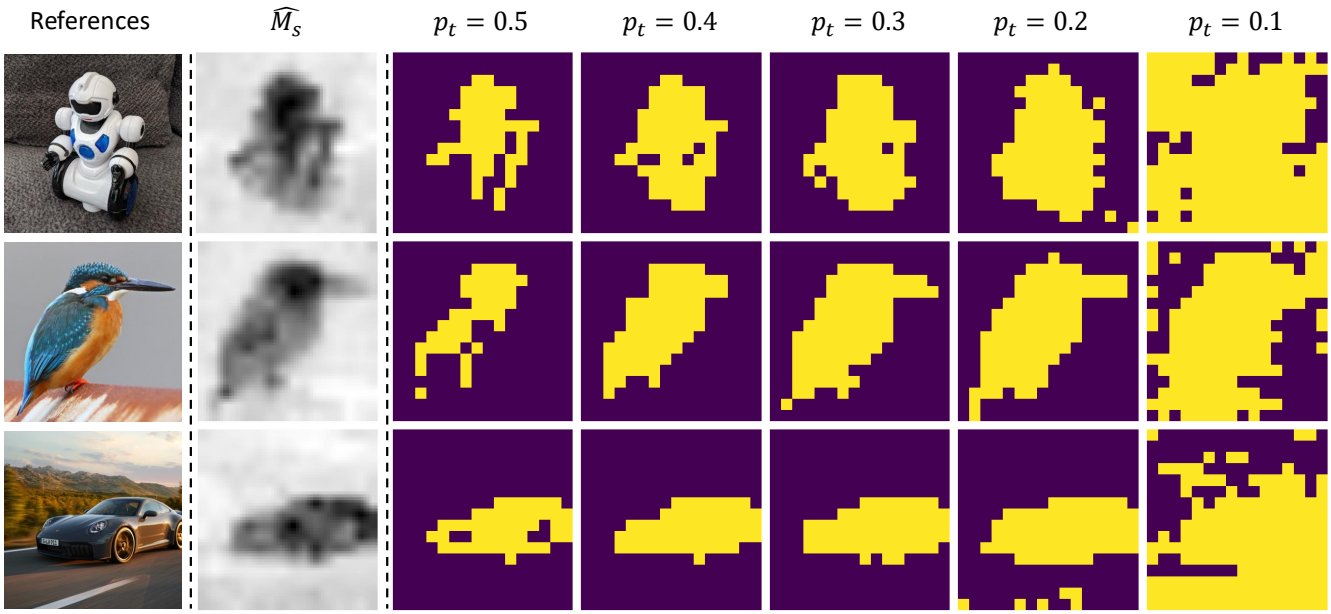

Fig. 2: **Comparing different values of p_t**

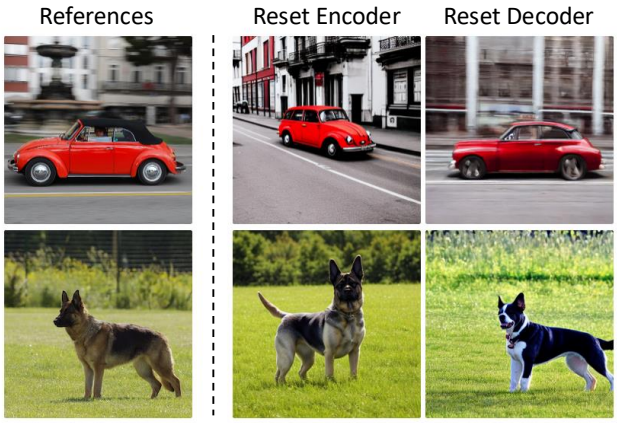

Fig. 3: **Understanding Encoder and Decoder Learning.** We fine-tune the entire U-Net on a single reference image. Subsequently, we reset either the encoder or the decoder by replacing their parameters with the pre-trained ones. We observed that resetting the encoder preserves the object's appearance but causes a loss of layout, whereas resetting the decoder preserves the layout but loses fine-grain appearances.

B. User Study

We conduct a user study comparing *Stencil* with the previous state-of-the-art, *DreamBooth*. Using the *DreamBench* dataset, we evaluate all live subjects across a set of various prompts. Each image is evaluated on subject consistency and text-to-image alignment.

```
Subject Consistency: Inspect the subject of
    ↪ the reference image. Select which of
    ↪ the images best reproduces the
    ↪ identity of the reference subject.


Text-to-Image Alignment: Select which of the
    ↪ images best follows the prompt [target
    ↪  prompt].

If you are unsure, or believe that the images
    ↪  equally follow the prompt, select '
    ↪ Undecided'.
```

C. Discussions

C.1. Ethical Concerns

A primary concern is the potential misuse of deepfakes,which can harm reputations and spread misinformation. To mitigate these risks, greater transparency around the use and origin of AI-generated content is essential.

C.2. Applications

Below, we present a representative (non-exhaustive) list of applications enabled by Stencil in Fig. 4, 5, 6, 7, 8, 9

D. Additional Qualitative Results

We provide additional qualitative results in Fig. 10, 11

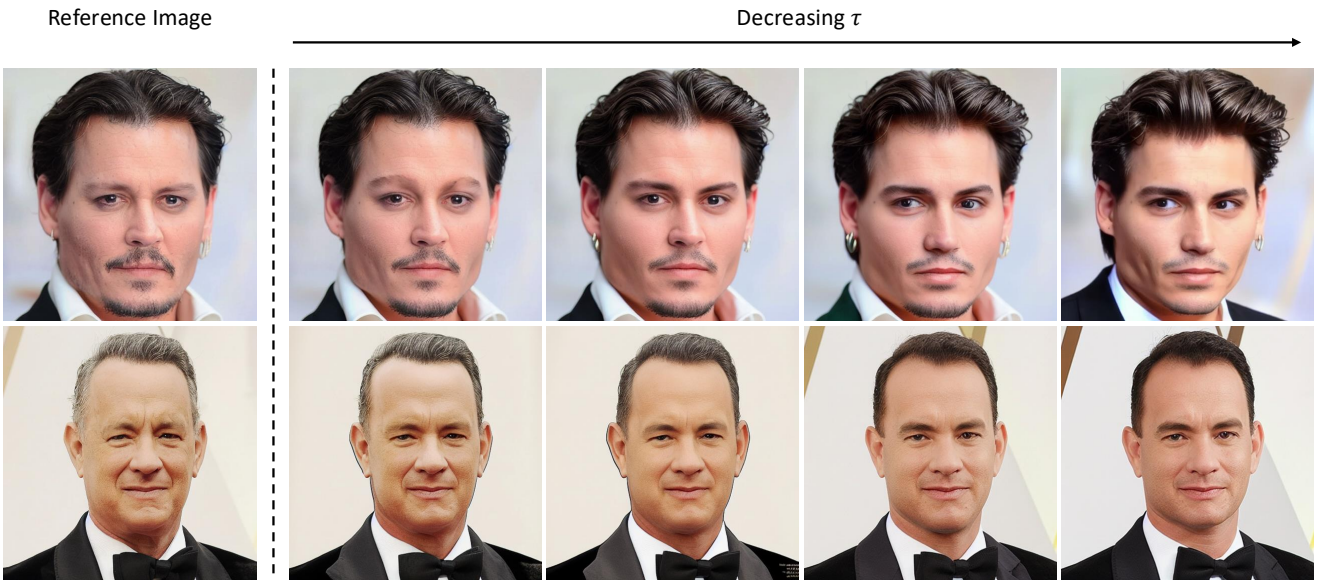

Fig. 4: **Age Progression/Regression.** We can fine-tune on images of the subject's younger self and, given a current image, interpolate observed age by adjusting the parameter τ. Notably, the generated younger versions exhibit a strong resemblance to how the subjects actually appeared in their youth.

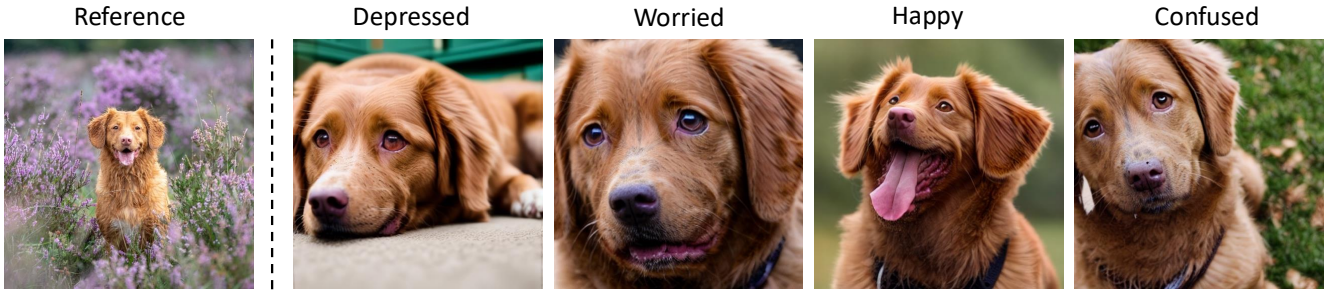

Fig. 5: **Expression Manipulation.** Stencil supports the generation of a diverse range of expressions of the subject while maintaining high subject fidelity using the prompt "A [emotion] [subject token]" at inference.

Fig. 6: **Accessorization.** We can generate the subject in various accessories by using the prompt "A [subject token] wearing [accessory]." at inference.

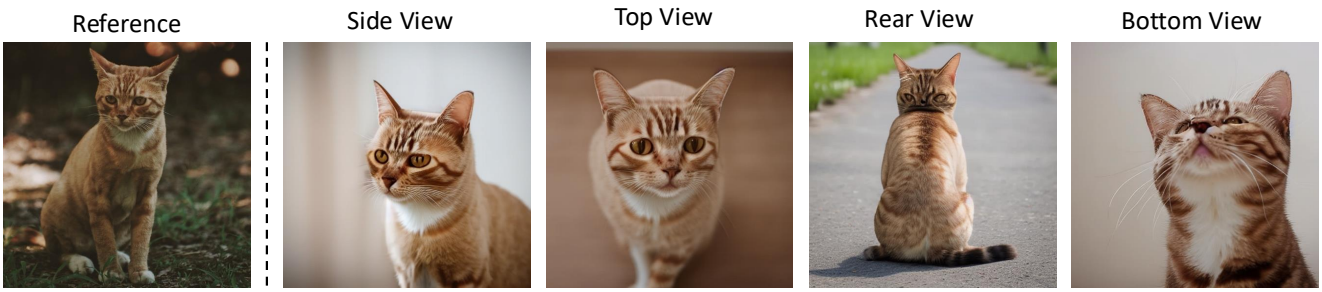

Fig. 7: **Perspective-conditioned Generation.** We can generate diverse images of the subject in different points-of-view, previously unseen in the reference images, using the prompt ”A [subject token] seen from [angle]”.

Fig. 8: **Pose Editing.** Stencil can generate diverse unseen poses of the subject that is beyond the generation capabilities of the small base model. We can achieve this using the prompt “A [subject token] [pose]”.

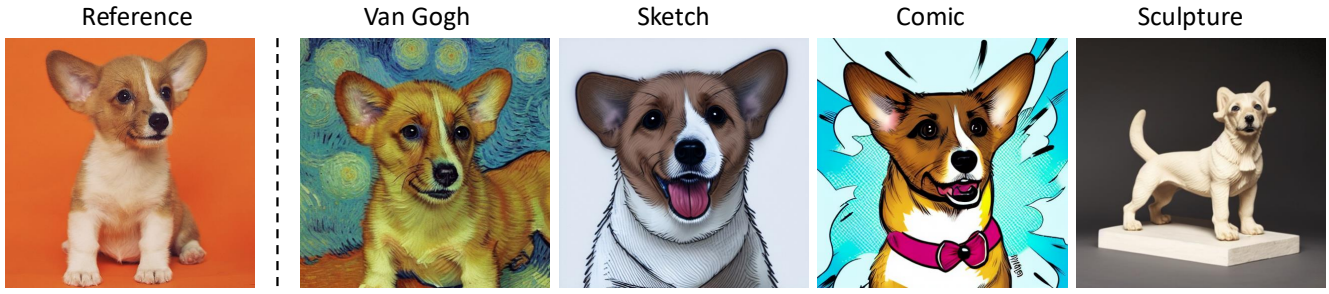

Fig. 9: **Style Transfer.** Stencil enables the seamless transfer of the subject to various artistic mediums, such as paintings and sculptures while maintaining key visual characteristics using the prompt “A [subject token] in [artistic style]”.

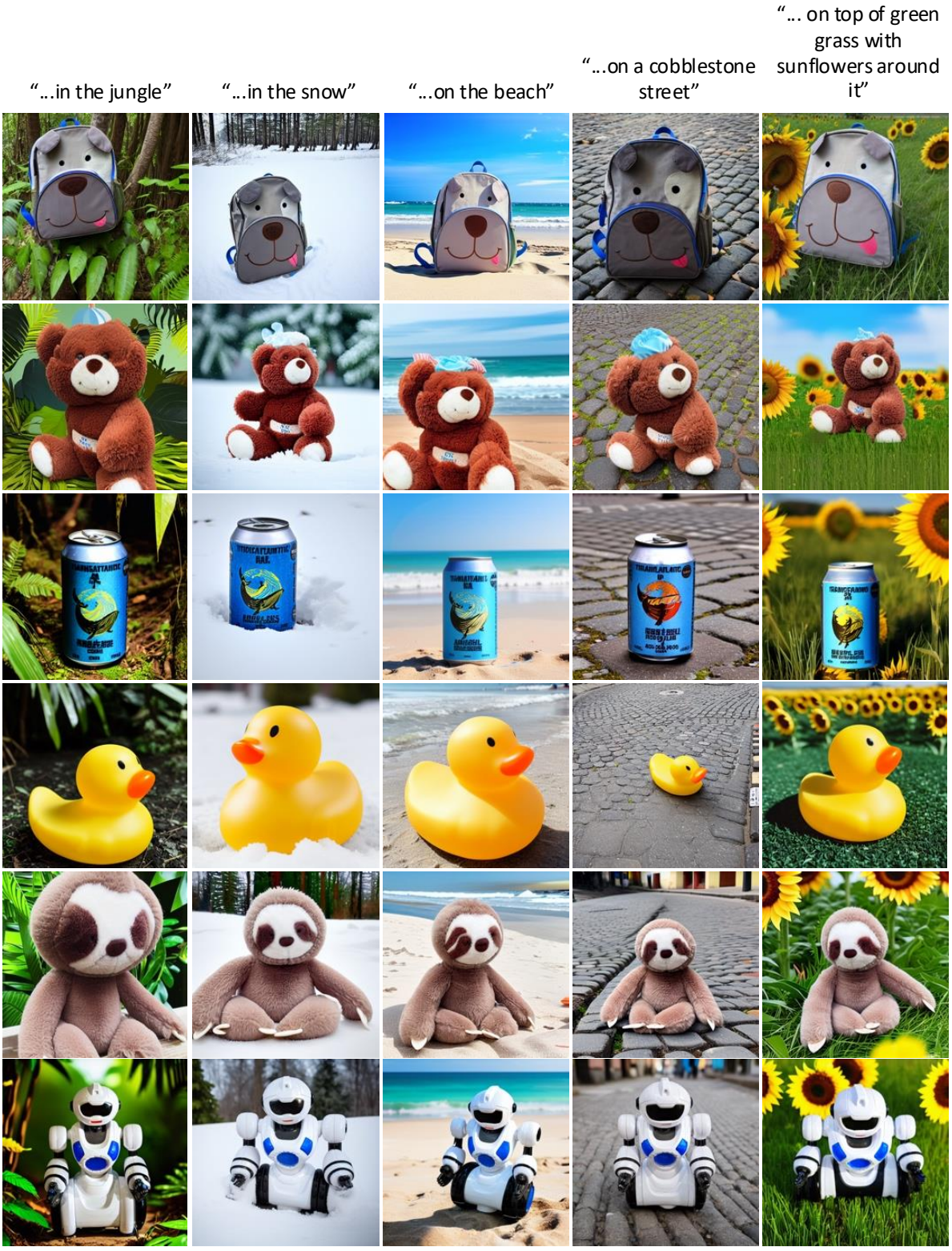

Fig. 10: **DreamBench Qualitative Results Part 1**.

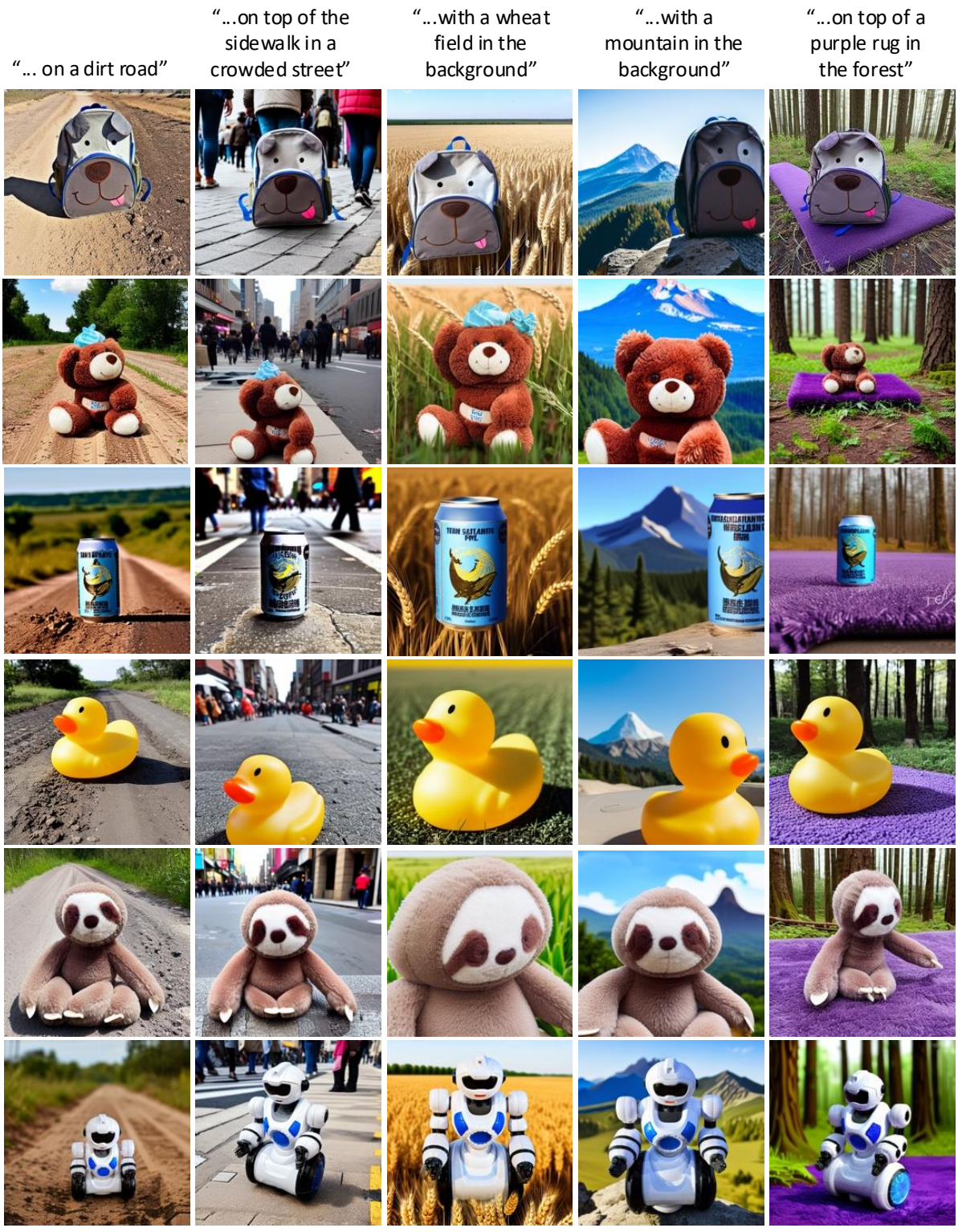

Fig. 11: **DreamBench Qualitative Results Part 2**.