ShapeWords: Guiding Text-to-Image Synthesis with 3D Shape-Aware Prompts

Supplementary Material

6. Implementation details

6.1. Data generation details

For depth images, we used the inverted ShapeNet data provided by the ULIP authors [47]. As stated in the main paper, for each depth image, we applied a randomly selected prompt from a set of 13,716 prompts for ControlNet conditioning. The ControlNet-based generation was done for 50 steps with control strength of 2. To promote adherence to shape geometry and reduce appearance biases during training, we additionally used the Stable Diffusion 2.1 inpainting model [38] for 50 steps to modify the backgrounds while preserving the foreground objects. The inpainting strength was set to 0.5.

6.2. SDS weighting function

For the SDS optimization loss of Eq. 6, we use the weighting function $W(t)$, proposed by DreamTime [18], that enhances training stability:

$$W(t) = \frac{1}{Z}\sqrt{\frac{1-\hat{\alpha}_t}{\hat{\alpha}_t}}\,\exp\left(-\frac{(t-m)^2}{2s^2}\right),$$

where m and s are hyperparameters controlling the weight distribution at each time step; $\hat{\alpha}_t$ is the noise scale for step t; and Z is a normalization constant ensuring that the weights sum to one over all timesteps. We set $m = 500$ and $s = 250$, which provide a good balance between high-frequency details (fine geometry) and low-frequency details (coarse geometry).

6.3. Training

We trained the model for 55 epochs on four NVIDIA A5000 GPUs with batch size 24 per GPU. The learning rate was set to 0.0005 with 1,000 warm-up steps to help stabilizing the training process. Similarly to textual inversion pipelines, we randomly crop and resize the training images to prevent overfitting of the model to spatial positions. The maximum scale of the crop was set to 0.8.

During training, the guidance prompt delta $\delta\mathbf{T}$ is applied to *all 77 word embeddings* (padding was set to max sequence length). We empirically found that this strategy during training helps the model to better generalize compared to adding the guidance delta to the object and EOS tokens only. We suspect that the usage of deltas on all token embeddings during training helps the model to diffuse training appearance biases across all tokens, which in turn reduces the overall appearance biases distilled in the object and EOS tokens.

Figure 10. **Token replacement strategies.** We qualitatively compare the following strategies for guidance: adding the prompt delta $\delta\mathbf{T}$ to all prompt embeddings; adding it to only the object word embedding; adding it only to EOS token embedding; adding it to both EOS and object token embeddings (as done in the main paper). Prompts are: 'a charcoal drawing of **chair**' (top row), 'Hieronymus Bosch's painting of a **chair**' (middle row), 'a **chair** under a tree' (bottom row). Target shapes are shown on the left. Compared to modifying the object & EOS tokens, the "all tokens" strategy produces over-smoothed images; the "object only token" strategy struggles to incorporate stylistic cues from the text into geometry; and the "EOS token" strategy struggles with preserving the target shape geometry.

Figure 11. **Failure cases – shape adherence.** Our model struggles to generalize to shapes with complex fine-grained geometries (e.g. a lot of thin parts or lot of holes). Prompts for the shapes are: 'a **chair**' (first two shapes); 'a **lamp**' (last three shapes). Target shapes are shown on the top.

7. Running times

We note that ShapeWords and ControlNet-based baselines rely on the same Stable Diffusion model (Stable Diffusion 2.1 base) and have similar computation requirements at test time: given a text prompt or/and depth image, it takes a few seconds to generate an image with 100 diffusion steps on a single GPU: 6.79s for ControlNet; and 5.00s for Stable Diffusion 2.1 with ShapeWords. The forward pass of the Shape2CLIP module takes 0.003s with pre-computed

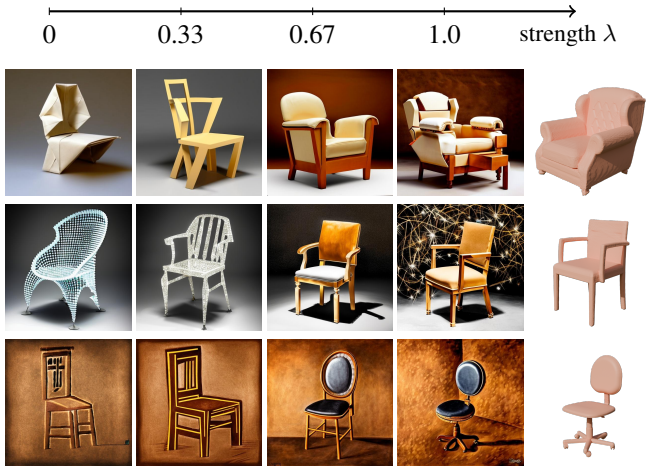

Figure 12. **Failure cases – prompt adherence.** Our model struggles to generalize to out-of-distribution prompts that require local adjustments of surface geometry. Prompts are: 'an origami of a chair' (top); 'a diamond sculpture of a chair' (middle); 'a hieroglyph of a chair' (bottom). Target shapes are shown on the right.

PointBERT embeddings. All measurements were done on single A40 GPU with batch size 1 and based on an average across 20 runs. For details on PointBERT computational costs, we refer to [49].

8. Token replacement strategies

We qualitatively compare token replacement strategies in Figure 10 at test time. Adding the guidance prompt delta $\delta\mathbf{T}$ to all tokens in the prompt yields overly smooth images that do not adher well to stylistic cues provided in the text or the target geometry. Adding $\delta\mathbf{T}$ only to object token without addition to the EOS token results in good geometry but still poor adherence to the stylistic cues in the prompt. Conversely, modifying only the EOS token results in good stylistic adherence but poor geometry. The strategy described in the main text, which is to add $\delta\mathbf{T}$ to both the object and EOS tokens, yields the best balance of textual and target shape adherence.

9. Additional quantitative results

We provide additional quantitative results in Table 2. Our model consistently outperforms ControlNet-Stop@K variants in terms of aesthetic score. In terms of CLIP score, we outperform all ControlNet-Stop@K variants, except for ControlNet-Stop@30 that matches the CLIP score of our method. Yet, as we discussed in our experiments in the "simple prompts dataset" as well as our perceptual user study in the "compositional prompts dataset", this variant severely underperforms in terms of shape adherence compared to our method. According to our user study, it also underperforms with respect to textual cue matching, when this is evaluated perceptually.

Model	**Aes.** ↑	**CLIP** ↑
ControlNet	5.24	26.9
CNet-Stop@20	5.15	30.3
CNet-Stop@30	5.18	**31.5**
CNet-Stop@40	5.15	30.3
CNet-Stop@60	5.20	28.3
CNet-Stop@80	5.17	27.5
ShapeWords	**5.45**	**31.5**

Table 2. **Evaluation results on compositional prompts.** Taking into account both the Aesthetics score and the CLIP score (scaled by 100), our method outperforms ControlNet variants in the challenging compositional setting. Even if the CNet-Stop@30 variant matches the CLIP score of our method, it still severely underperforms in terms of shape adherence according to our user study and the rest of our experiments.

10. Additional qualitative results

We provide additional qualitative results for shape and prompt adherence in Figures 13 and 14, respectively.

11. Failure cases

We observed that the failure cases for our model fall in two modes. First, it struggles with capturing details of challenging fine-grained geometry (Figure 11). In such cases, ShapeWords correctly captures coarse shape structure but struggles to reproduce fine geometric details. Our hypothesis is that the geometric precision of ShapeWords is likely to be bound by the image resolution of OpenCLIP model (ViT-H/14, 224px) which we used to train ShapeWords, and the ability of PointBert to capture such fine-scale geometric details. Training ShapeWords with variants of CLIP of higher resolution might yield better geometric precision – we consider that this is a promising direction for future work.

Second, our model struggles to generalize to largely out-of-distribution text prompts. We illustrate this issue in Figure 12. For example, the prompt 'an origami of a chair' requires both adjustment of texture and local geometry. Our model struggles to do both, especially for high values of guidance strength. We think this issue arises from a combination of two factors: a) our set of prompts is biased towards smoother appearances (e.g. 'photo', 'sketch', 'illustration'), b) our supervisory images come from ControlNet that also tends to produce smooth surfaces following depth maps. However, results for intermediate guidance strength suggest that our model can still generalize to such prompts to some extent. We suspect that this issue could potentially be alleviated by using more diverse training data.

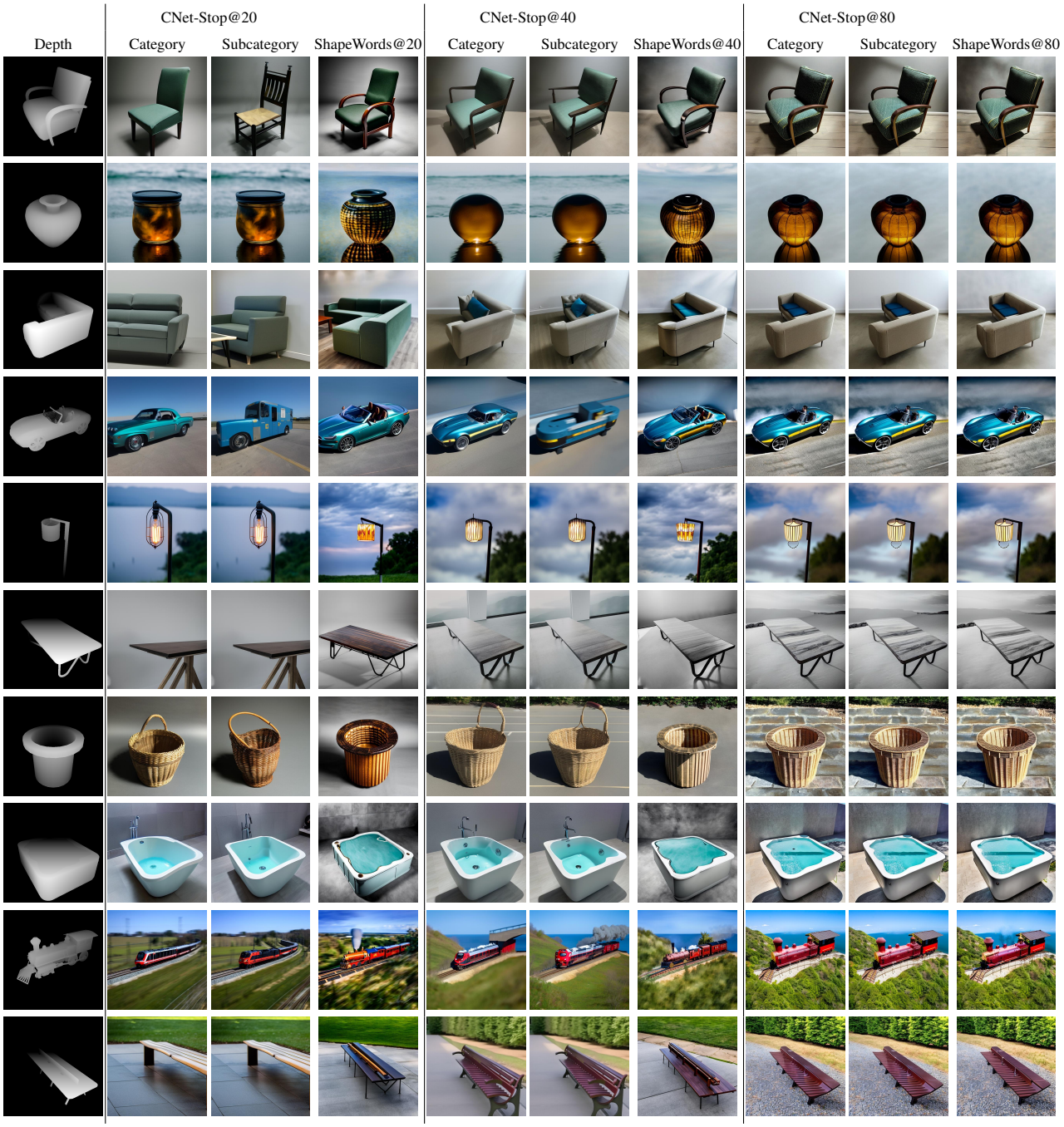

Figure 13. **Shape adherence – additional examples.** ShapeWords@20 produces shapes that are significantly more consistent with target shape geometry compared to the CNet-Stop@20 (conditioned either on category or subcategory prompts). ShapeWords@40 seems still more shape-adhering than CNet-Stop@40. In the setting of ShapeWords@80 and CNet-Stop@80, which both become more over-constrained by depth, differences become less noticeable.

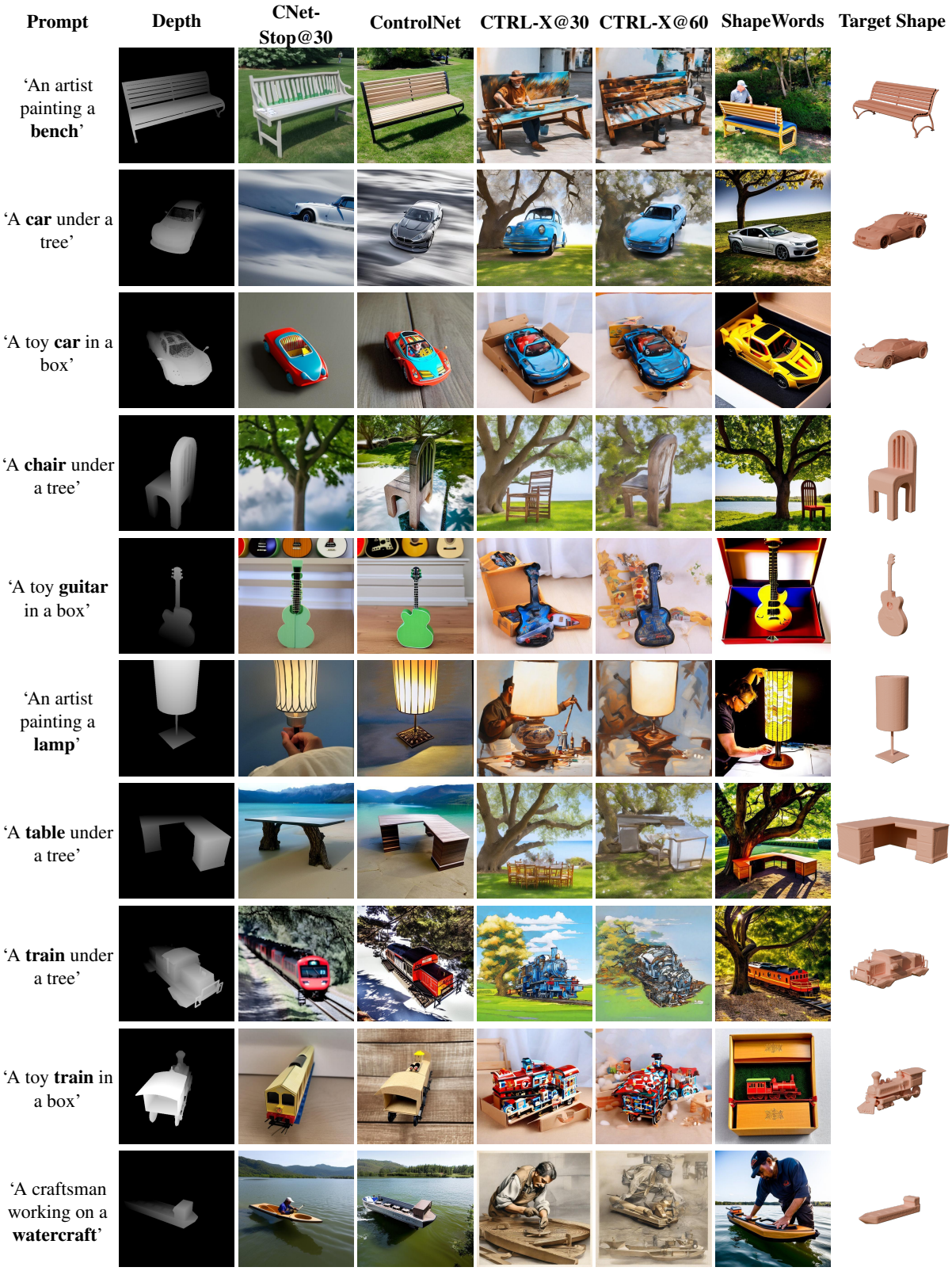

Figure 14. **Generalization to compositional prompts – additional examples.** Baselines that heavily rely on input depth maps (e.g. ControlNet and Ctrl-X@60) appear to be over-constrained and ignore prompt composition. In contrast, baselines that are under-constrained by the input depth (e.g. CNet-Stop@30 or Ctrl-X@30) stray too much from target shape geometry. ShapeWords achieves much better generalization to compositional prompts, while still demonstrating strong adherence to the target shape.