Zero-Shot Text-Guided Object Generation with Dream Fields Supplementary Material

Anonymous CVPR submission

Paper ID 11674

1. Supplementary video

An explanatory supplementary video is included in the zip file, and available at https://dreamfieldspaper.github.io. The video includes 360° renderings where the camera is moved around the object, as well as associated depth maps.

2. Object Centric COCO dataset

153 test set prompts and 74 development set prompts from the Object Centric COCO dataset are included in text files in the supplementary zip. Some additional prompts are used for qualitative results, and are included in the main paper alongside figures.

3. Hyperparameters and training setup

Positional encoding Our Fourier feature positional encodings use $L = 8$ frequency levels, while novel view synthesis applications with image supervision commonly use $L = 10$ to fit high-frequency details in photographs. Low-frequency ablations in Table 1 use $L = 6$, which can improve convergence in the absence of our other geometric priors.

Rendering Scenes are bounded to a cube with side length 2. The camera is sampled at a fixed radius of 4 units from the center of the cubical scene bounds and an elevation of 30° above the equator. Near and far planes are set at $4 \pm \sqrt{3}$ units from the camera based on the minimum and maximum possible distance to the corners of the cube. During training, we sample 192 points along each ray, spaced uniformly and jittered with uniform noise. Rendered 168^2 views are cropped to 154^2 and upsampled to CLIP's input resolution for scenes where we compute qualitative metrics or 252^2 views are cropped to 224^2 for certain higher-quality visualizations. Crop sizes are selected to cover about 80% of the image area. At test time, we sample 512 points along the rays and render at a higher resolution equal to CLIP's input size of 224^2 or LiT's input size of 288^2 for computing R-Precision and 400^2 for visualizations.

Optimization MLP parameters are optimized with Adam with $\epsilon = 10^{-5}$. Learning rate warms up exponentially from 10^{-5} to 10^{-4} over 1500 iterations, then is held constant. The camera origin is separately tracked with an exponential moving average with decay rate 0.999 of the center of mass of rendered density.

Hyperparameter selection Hyperparameters are manually tuned for visual quality on a development set of 74 object-centric COCO captions distinct from the test set reported in the paper. Most tuning is done on a smaller subset of 20 of the 74 captions, and hyperparameters are shared across all scenes.

Hardware Optimization is done on 8 preemptible TPU cores, and 10K iterations takes approximately 1 hour 12 minutes. This means each Dream Field costs approximately $3 to generate on Google Cloud, which is economical for applications. Training is bottlenecked by MLP inference and backpropagation during volumetric rendering, not CLIP.

4. Impact of optimization time

Dream Fields can overfit to the aligned image-text representation used for optimization. Figure 1 shows the training losses, $\mathcal{L}_{\text{CLIP}}$ and mean transmittance $\text{mean}(T(\theta, \mathbf{p}))$, as well as the validation R-Precision according to a different contrastive image-text model. Validation renderings are also done at a held-out elevation angle. Training loss continues to improve over long optimization trajectories, up to 10× longer than reported in the main paper. However, validation retrieval accuracy declines after 5-10K iterations.

Qualitatively, additional details and hyper-realistic effects are added over the course of long runs. Some details are not realistic, like floating text related to the typographic attacks identified in [1].

More augmentations may help further regularize the optimization. These include more aggressive 2D image augmentations such as smaller random crops, and more 3D

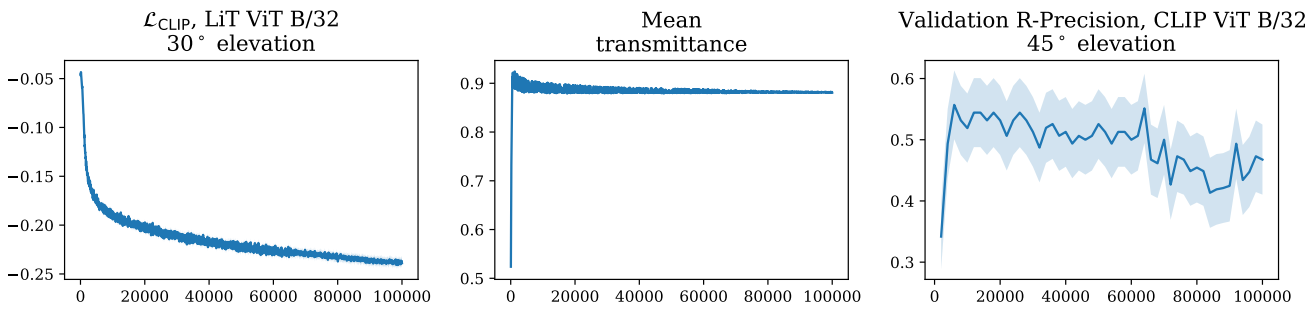

Figure 1. Long-run training and validation curves averaged over Object Centric COCO captions. Transmittance remains close to the target τ throughout training. Dream Fields overfit to the image-text representations used for optimization, so in quantitative experiments, we stop training at 10K iterations. The standard error of the mean is shaded.

data augmentations including varying focal length, varying distance from the subject and varying elevation.

5. Signed distance field parameterizations

In early experiments, we learned scene density σ with the VolSDF parameterization [2] $\sigma(\mathbf{x}) = \alpha\Phi_\beta(-d_\Omega(\mathbf{x}))$ where $d_\Omega(\mathbf{x})$ is a signed distance function implicitly defining the object surface and Φ is the CDF of the Laplace distribution. This allows normal vector prediction with autodifferentiation and could improve the quality of the surface extracted from the radiance field. Dream Fields successfully train with this alternate parameterization and produce visually compelling objects, but the additional Eikonal loss was occasionally hard to balance and tune. Alterate 3D representations are an interesting avenue for future work.

References

[1] Gabriel Goh, Nick Cammarata †, Chelsea Voss †, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. Multimodal neurons in artificial neural networks. *Distill*, 2021. https://distill.pub/2021/multimodal-neurons. 1

[2] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *arXiv:2106.12052*, 2021. 2