Compositional Visual Generation with Composable Diffusion Models Supplementary Material

Nan Liu[1⋆], Shuang Li[2⋆], Yilun Du[2⋆]
Antonio Torralba[2], and Joshua B. Tenenbaum[2]

[1] University of Illinois Urbana-Champaign
[2] Massachusetts Institute of Technology
nanliu4@illinois.edu, {lishuang,yilundu,torralba,jbt}@mit.edu

In this appendix, we first provide additional results in appendix A. We then show the details of training classifiers in appendix B. In appendix C and appendix D, we show more details of our approach and baselines, respectively. Finally, we provide the implementation details in appendix E.

A Additional Results

In this section, we first show results of composing human facial attributes in appendix A.1 as we described in the main paper Section 6. We then show more qualitative results in appendix A.2.

A.1 Composing Human Facial Attributes

Qualitative results. We compare the proposed method and baselines on compositing facial attributes in Figure 1. We find that *LACE* and *StyleGAN2* can generate high-fidelity images, but the generated images do not match the given label. For example, *StyleGAN2* generates humans without wearing glasses when the input labels contain "glasses". *LACE* generates males sometimes when the input is "NOT Male". The image quality of *EBM* is much worse than other methods. In contrast, our method can generate high-fidelity images, containing all the attributes in the input label.

Quantitative results. The results of our method and baselines on three test settings are shown in Table 1. Our method is comparable with the best baseline on each evaluation metric.

A.2 More Qualitative Results

⋆ indicates equal contribution.
Correspondence to: Shuang Li <lishuang@mit.edu>, Yilun Du <yilundu@mit.edu>

Fig. 1: **Composing Facial Attributes.** Image generation results on the FFHQ dataset. Our model is trained to generate images conditioned on a single human facial attribute, but during inference, our model can recursively compose multiple facial attributes using the proposed compositional operators. The baselines either fail to compose attributes (StyleGAN2 and LACE) or generate low-quality images (EBM).

We provide more qualitative results of the proposed method on composing concepts using the conjunction operator. Figure 3, 4, 5 and 6 shows the results of compositing language descriptions. Figure 7 shows additional results on compositing objects on the CLEVR dataset. Our approach can reliably generate images conditioned on multiple concepts, even for combinations outside the training distribution

We further show results of compositing facial attributes on the FFHQ dataset in Figure 8. Our model is trained to generate images conditioned on a single human facial attribute, but it can compose multiple attributes during inference without further training using the conjunction and negation compositional operators. As shown in the fifth row of Figure 8, our model can compose *Not Male* and *Glasses* and generate images with females wearing glasses. The proposed compositional operators allow our model to compose facial attributes recursively.

Fig. 2: Our method (composing multiple sentences) generates different styles of images compare to *GLIDE* (directly encodes the descriptions as a single long sentence).

Interesting cases. As shown in Figure 2, we find that our method, which combines multiple textual descriptions, can generate different styles of images compared to *GLIDE*, which directly encodes the descriptions as a single long sen-

Table 1: Image generation results on FFHQ. The binary classification accuracy (Acc) and FID are reported. Our method achieves comparable results with the best baselines on three test settings.

Models	1 Component		2 Components		3 Components	
	Acc (%) ↑	FID ↓	Acc (%) ↑	FID ↓	Acc (%) ↑	FID ↓
EBM [1]	98.74	89.95	93.10	99.64	30.01	335.70
StyleGAN2 [5]	58.90	**18.04**	30.68	18.06	16.96	18.06
LACE [11]	97.60	28.21	**95.66**	36.23	**80.88**	34.64
GLIDE [9]	98.66	20.30	48.68	22.69	27.24	21.98
Ours	**99.26**	18.72	92.68	**17.22**	68.86	**16.95**

tence. Prompted with "a dog" and "the sky", our method generates a dog-shaped cloud, whereas *GLIDE* generates a dog under the sky from the prompt "a dog and the sky".

B Details of Binary Classifiers

We provide more details of the binary classifiers in this section.

CLEVR. CLEVR dataset consists of 30,000 image-label pairs. We split the dataset into training and validation subsets. There are $24,000$ data pairs used for training and $6,000$ data pairs used for validation. We train a binary classifier to evaluate whether there is an object appearing at a particular position of an image. The classifier achieves 99.05% accuracy on the validation set, which is used to evaluate the quality of generated images.

Relational CLEVR. Relational CLEVR [7] contains $50,000$ images at 128×128 resolution. We split the dataset into $40,000$ training data and $10,000$ validation data. Then we train a binary classifier to evaluate whether an image contains an object relational description. The classifier achieves 99.80% accuracy on the validation set.

FFHQ. We use $30,000$ image-label pairs from CelebA-HQ [4] to train a classifier for FFHQ generated images. We split the dataset into training and validation subsets using $80:20$ ratio. We select three attributes (*i.e. smiling*, *glasses*, and *gender*) to evaluate the compositionality ability of our approach and baselines. We thus train three binary classifiers to evaluate the *smiling*, *glasses*, and *gender* concepts respectively. Our classifiers achieve 95.01%, 99.20% and 97.49% accuracy on the validation sets of *smiling*, *glasses*, and *gender*.

To further verify the reliability of results obtained by the classifiers, we add human evaluation results and find that our method still outperforms baselines. We generated 300 facial images using our method and one of the best baselines (LACE), respectively. Given a concept combination, *e.g. Smiling AND (NOT Male)*, each method generates an image conditioned on this combination. We asked workers to select which image matches the input concepts the best. At 62% of the time, the workers think the images generated by our method are better.

C Details of Our Approach

Training. Our approach is implemented based on the code from [10,9]. Ho *et al*. [3] introduce a technique to train a conditional and an unconditional diffusion model at the same time by masking some labels as null labels. During training, we utilize the same approach. We randomly replace 10% of training labels as the null labels in our training to estimate the unconditional score and otherwise use conditional labels.
Inference. To generate images, we compute the unconditional and conditional scores for each label and use the combined score to sample a less noisy image at each timestep. To generate FFHQ images, we first generate images at 64×64 resolution and then upsample the images to 256×256. For CLEVR images, we generate images at 128×128 resolution directly.
Label Encoding. On the FFHQ dataset, we use three attributes, including *smile*, *glasses* and *gender*. For the *smile* and *glasses* attributes, label 1 indicates that the image contains the attribute, and label 0 indicates its absence. For the *gender* attribute, label 0 indicates "male", while label 1 represents "female". We use the embedding layer $nn.Embedding(7, d)$ to encode the attribute labels, including 6 attribute labels and 1 null class label. The labels are encoded as a d-dimension feature vector, which is then fused with the embedding of the iteration step t and image x_t. The fused features are sent to the U-Net [12] during training.

On the CLEVR dataset, we encode the (x, y) coordinates using a linear layer $nn.Linear(2, d)$, where 2 is the dimension of the (x, y) coordinates and d is the dimension of the hidden feature. The coordinates embedding is then fused with the embedding of the iteration step t and image x_t, which are further sent to the U-Net [12] during training.

D Details of Baselines

Energy-based models (EBMs). We train energy-based models using the codebase from [2], where we encode discrete labels and continuous labels using an embedding layer and a linear layer, respectively. We use the inference code from [1] to compose multiple concepts.
StyleGAN2. We train an unconditional StyleGAN2 on CLEVR, while we use an existing StyleGAN2 model trained on FFHQ. For training, we use the "config-f" setting provided by [5]. To enable image generation conditioned on multiple concepts, we train a binary classifier on each dataset. During inference, we optimize the underlying latent code to minimize each loss from the classifier conditioned on each individual label.
LACE. LACE [11] trains classifiers for image generation by using sampled images from StyleGAN2 and labels provided by the neural network. For CLEVR dataset, we firstly generate $10,000$ images using the same StyleGAN2 model that was trained on CLEVR in Section D. Then we modify the code to train a position annotator using a DenseNet model provided by LACE to label the positions of

generated images. Lastly, we train a classifier conditioned on coordinates using their provided script. For FFHQ, we use their off-the-shelf pre-trained model for comparison. To enable image generation, we utilize their inference scripts.

GLIDE. We use the released GLIDE [9] model in our experiments. We develop Composed GLIDE (Ours), a version of GLIDE that utilizes our compositional operators to combine textual descriptions, without further training. We compare it to the original GLIDE, which directly encodes the descriptions as a single long sentence. [9] also released a upsample model to upsample the generated images to a resolution of 256×256. We use the upsample model for both the GLIDE and Composed GLIDE (Ours).

E Implementation Details

EBMs. In our experiments, we use the same setting to train models on different datasets. We use the Adam optimizer [6] with a learning rate of 10^{-4}. For MCMC sampling, we use a step size of 300 and 80 iterations. On each dataset, the model is trained for two days on a single Tesla 32GB GPU.

StyleGAN2. We train the StyleGAN2 model for 2 days on CLEVR using a single Tesla 32GB GPU. It takes 2 hours to train binary classifiers for each dataset. We use the Adam optimizer [6] with $\beta_1 = 0$, $\beta_2 = 0.99$, and $\epsilon = 10^{-8}$ to train the StyleGAN2 model (more details can be found in the codebase from [5]). We use the Adam optimizer with $\beta_1 = 0$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$ to train the classifiers. We use the pre-trained model provided by [5] on the FFHQ dataset.

LACE. LACE uses the pre-trained model provided by [5] on the FFHQ dataset as well. For CLEVR, we use the same StyleGAN2 model as described in Section E. It takes less than 10 minutes to train the classifier on each dataset using a single Tesla 32GB GPU.

Our Approach. To train diffusion models on both CLEVR and FFHQ, we use $1,000$ diffusion steps, and the cosine noise schedule. We use the AdamW optimizer [8] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. We train the diffusion models on CLEVR and FFHQ for 7 days ($750,000$ iterations) and 2 days ($250,000$ iterations), respectively, using a single Tesla 32GB GPU.

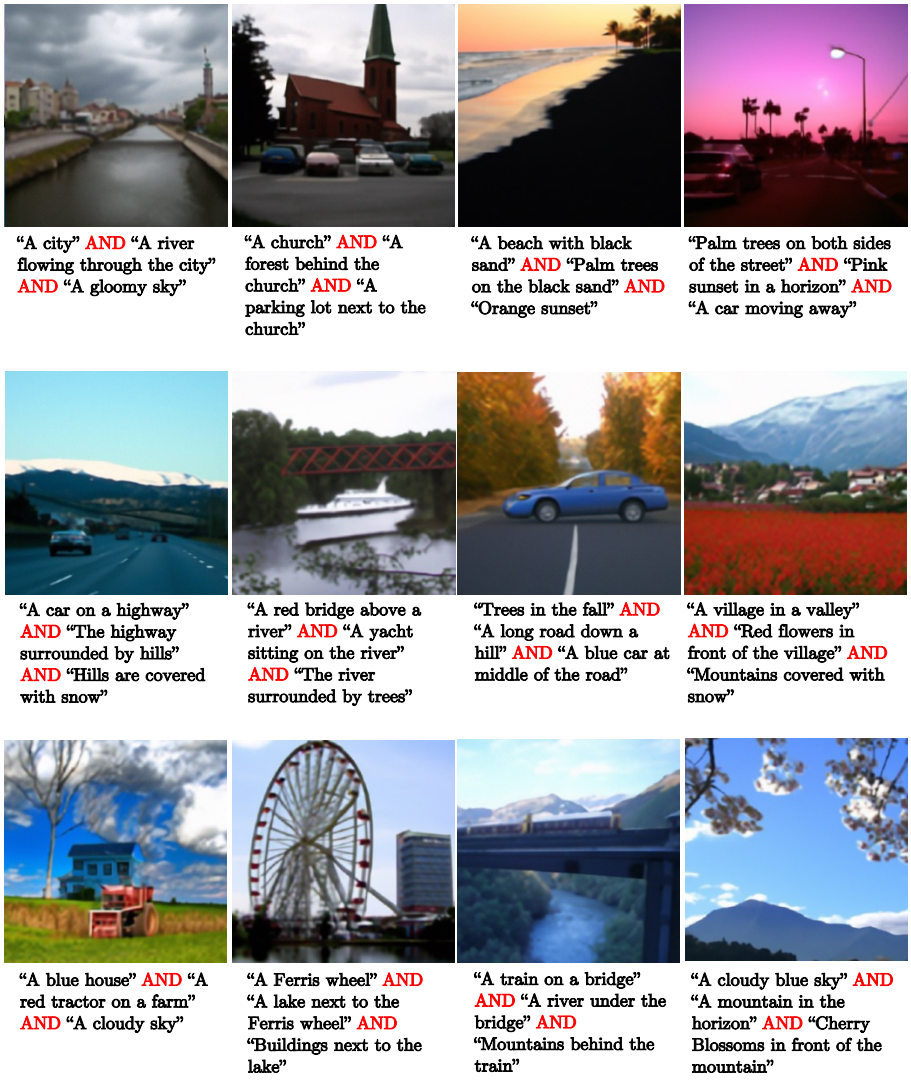
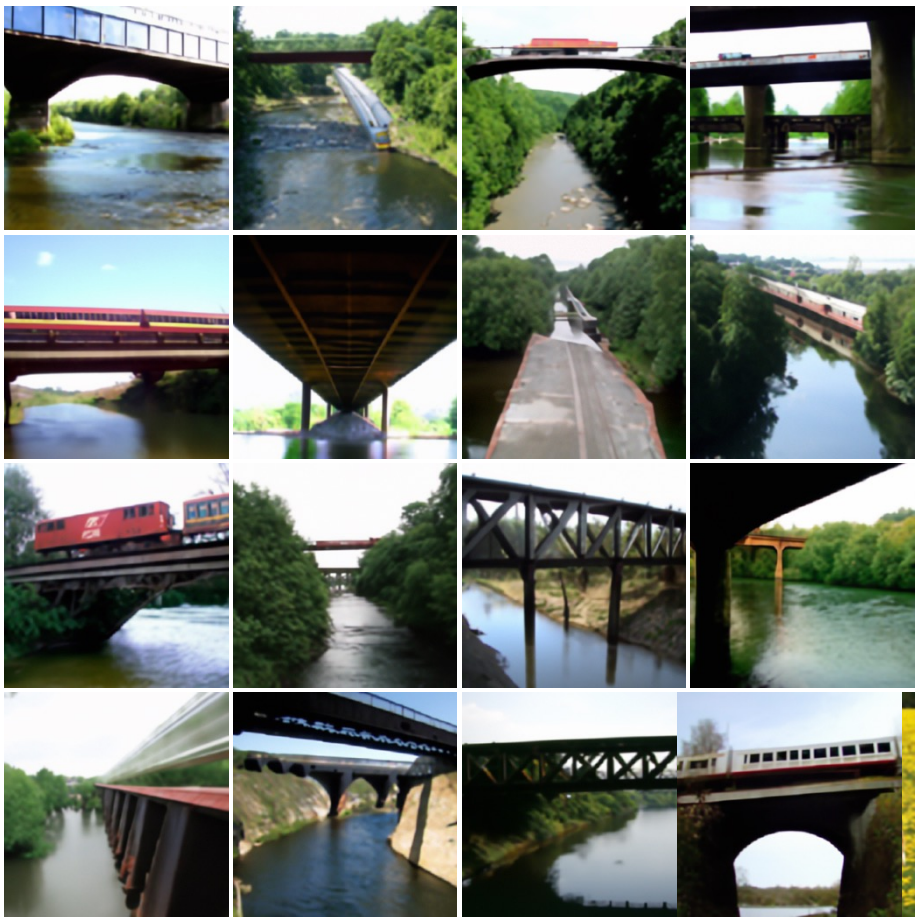


Fig. 3: **Composing Language Descriptions**. We provide more qualitative results of *Composed GLIDE (Ours)*, a version of GLIDE [9] that utilizes our compositional operators to combine textual descriptions, without further training.

"A river leading into mountains" AND "red trees on the side"

Fig. 4: **Composing Language Descriptions**. Images generated by our method, *Composed GLIDE (Ours)*.

Fig. 5: **Composing Language Descriptions**. Images generated by our method, *Composed GLIDE (Ours)*.



"A train on a bridge" AND "A river under the bridge"

Fig. 6: **Composing Language Descriptions**. Images generated by our method, *Composed GLIDE (Ours)*.

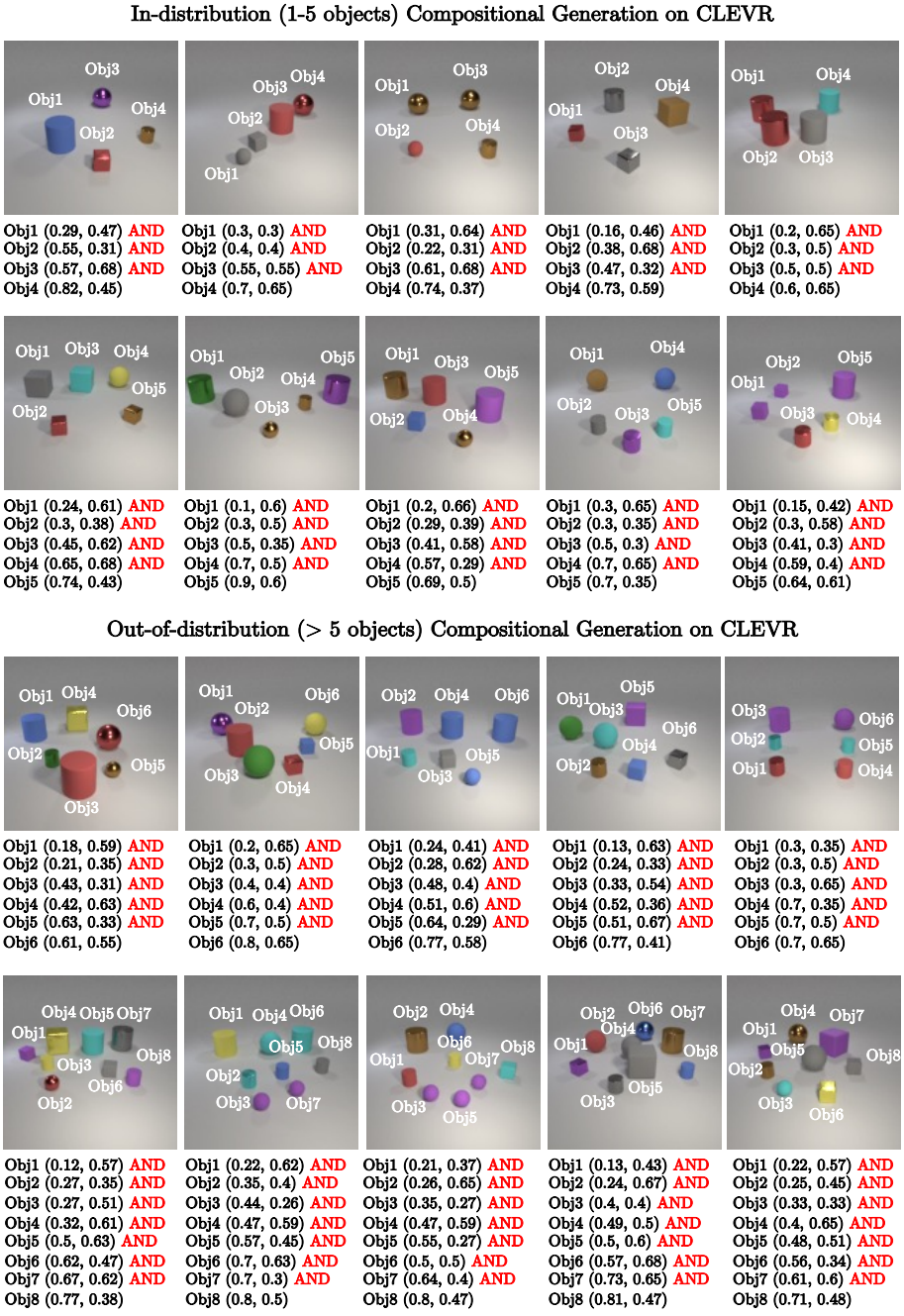


Fig. 7: **Composing Objects.** During inference, our model can generate images that contain multiple objects by composing their probability distributions using the conjunction operator. Note that the training set only contains images with fewer than 5 objects, but our model can compose more than 5 objects during inference.

Fig. 8: **Composing Facial Attributes.** During inference, our model can generate images that contain multiple attributes by composing their probability distributions using the negation and conjunction operators.

References

1. Du, Y., Li, S., Mordatch, I.: Compositional visual generation with energy based models. Advances in Neural Information Processing Systems **33**, 6637–6647 (2020)
2. Du, Y., Li, S., Tenenbaum, J., Mordatch, I.: Improved contrastive divergence training of energy based models. arXiv preprint arXiv:2012.01316 (2020)
3. Ho, J., Salimans, T.: Classifier-free diffusion guidance. In: NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications (2021)
4. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196 (2017)
5. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8110–8119 (2020)
6. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
7. Liu, N., Li, S., Du, Y., Tenenbaum, J., Torralba, A.: Learning to compose visual relations. Advances in Neural Information Processing Systems **34** (2021)
8. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
9. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021)
10. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: International Conference on Machine Learning. pp. 8162–8171. PMLR (2021)
11. Nie, W., Vahdat, A., Anandkumar, A.: Controllable and compositional generation with latent-space energy-based models. Advances in Neural Information Processing Systems **34** (2021)
12. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)