Compositional Visual Generation with Composable Diffusion Models

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

Abstract. Large text-guided diffusion models, such as DALLE-2, are able to generate stunning photorealistic images given natural language descriptions. While such models are highly flexible, they struggle to understand the composition of certain concepts, such as confusing the attributes of different objects or relations between objects. In this paper, we propose an alternative structured approach for compositional generation using diffusion models. An image is generated by composing a set of diffusion models, with each of them modeling a certain component of the image. To do this, we interpret diffusion models as energy-based models in which the data distributions defined by the energy functions may be explicitly combined. The proposed method can generate scenes at test time that are substantially more complex than those seen in training, composing sentence descriptions, object relations, human facial attributes, and even generalizing to new combinations that are rarely seen in the real world. We further illustrate how our approach may be used to compose pre-trained text-guided diffusion models and generate photorealistic images containing all the details described in the input descriptions, including the binding of certain object attributes that have been shown difficult for DALLE-2. These results point to the effectiveness of the proposed method in promoting structured generalization for visual generation.

Keywords: Compositionality, Diffusion Models, Energy-based Models, Visual Generation

1 Introduction

Our understanding of the world is highly compositional in nature. We are able to rapidly understand new objects from their components, or compose words into complex sentences to describe the world states we encounter [24]. We are able

* indicates equal contribution.
Correspondence to: Shuang Li <lishuang@mit.edu>, Yilun Du <yilundu@mit.edu>
Webpage: https://energy-based-model.github.io/Compositional-Visual-Generation-with-Composable-Diffusion-Models/