Exploring Stable Diffusion Models through Category Theory for Image Embeddings

Anonymous ICCV submission

Paper ID 6361

Abstract

Stable diffusion models have shown impressive results in generating high-quality images from text prompts. However, these models are not fully compositional and can generate incorrect images for complex descriptions. In this paper, we propose a novel approach to study the latent space structures of diffusion models using the framework of category theory. Our hypothesis is that the lack of an isomorphism between the text category and the image category is the reason for the failure modes observed in existing models. To address this, we propose means to learn the structure embedded in the text category where the morphisms between two text prompt objects denote the degree to which the prompts are related. We investigate to what extent the structure learned from the text category is preserved in the image embeddings space. Specifically, we explore the preservation of compositionality and the existence of universal properties in image embeddings.

1. Introduction

Diffusion models like DALLE-2 and Stable diffusion are some of the popular text to image tools available now. While they have been succesfull in image generation, they have many failure cases in compositional generation such as missing objects, attribute leakage and interchangeable attributes [2]. Some of the efforts aiming at improving composition generation in diffusion models have been successful but not perfect [2]. Previous works have aimed at modifying architectures or modifying the image embeddings so as to improve compositionality. In this work, we aim to provide a common mathematical language using category theory [3] to reason about the latent space of these diffusion models. Our goals are two fold – first, we would like to identify those embeddings that generate composable images and differentiate from those that cannot, second, adjust/improve the embeddings inorder to generate composable images.

Related Works

Related Works What people have worked on? What have they done before? How is this work building on top of them? How is this work utilizing the related works?

2. Cateogory Theoretic Perspective on Latent Space Embedings

Provides us tools to (a) study the structure and (b) perturb or identify emebeddins that can preserve the structure.

3. Evaluation Benchmark

We approach the problem by identifying three categories underlying the diffusion models – the language (that contains prompt strings), the text (containing the embeddings generated as a consequence of BERT or CLIP models) and lastly the image (containing the image embeddings). These image embeddings when passed through a decoder reconstructs the desired images.

3.1. Primer on Category Theory

We define objects as the entities such as the text prompts or the embeddings in the latent space in each category. The morphisms that relate one object to another are defined using the "belong to" operation similar to [**?**]. That is, there exists a morphism or an edge between two objects if the former text prompt is contained in the later. A weight can be assigned to the morphism that can correspond to the jaccard similarity between the prompts.

3.2. Identifying non-composable embeddings

The yoneda lemma offers us tools to identify non-composability of the embeddings. Specifically, we have

Lemma 1. *If* $X \simeq Y$ *if and only if* $hom(X, -) \simeq hom(Y, -)$.

The above result also relates to the generalization theorem when in the context of foundation models, see [5].

The above lemma begs the question of whether we can use the Yoneda perspective to compare the embeddings in

the text space with those in the image space. This forms the central research question of this paper.

3.3. Defining the categories with commutative pre-orders

[**Vishnu**: Modelling and design choices from [1]]

4. Experiments

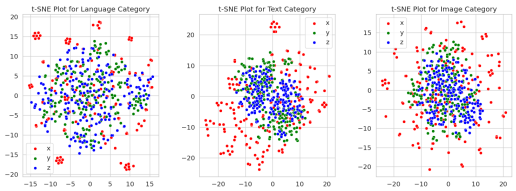
4.1. Setting

In order to better assess which classes of embeddings are work well and which do not, there is a need to systematically categorize the different prompts generating these embeddings. Hence, we define three classes of prompts and name them the X class, the Y class and the Z class. Each of these classes have increasing number of nouns with the X class containing single nouns, the Y class containing $2-3$ nouns and the Z class containing $3-4$ nouns. Our dataset consists of such triplets $X-Y-Z$ with about 20000 samples. More details are provided in the supplement.

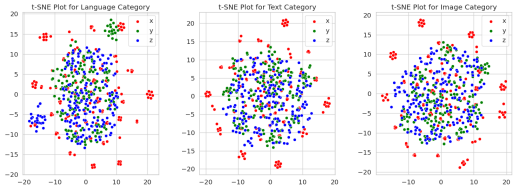
4.2. Evaluation Metrics

- Satisfaction of Transitivity Axioms
- Correlation with the CLIP scores

4.3. Structure better captured through Yoneda Embeddings



(a) Original Embeddings



(b) Yoneda Embeddings

Figure 1: Comparison of Original and Yoneda Embeddings We plot the t-SNE representations of the embeddings before and after enriching it with the Yoneda Lemma. A key takeaway is that the Yoneda embeddings are structurally similar in all of the three categories.

Table 1: Cosine Similarity of the Original and the Yoneda Embeddings

Original Embeddings

	Language	Text	Image
Language	1.0	0.545	0.547
Text	0.545	1.0	0.987
Image	0.547	0.987	1.0

Yoneda Embeddings

	Language	Text	Image
Language	1.0	0.940	0.910
Text	0.940	1.0	0.985
Image	0.910	0.985	1.0

4.4. Axiom Satisfaction of the Yoneda and the Original Embeddings

We test the satisfaction of the transitivity axiom in all of the three categories. The measure of the y-axis computes $C(x,y) * C(y,z) - C(x,z)$ 2. Here, x, y and z denote a triplet of prompts. It is desired that this measure is negative. Although both the original and the yoneda embeddings are all negative across the three categories, we observe that the variance of the measure is large in both the text and the image categories.

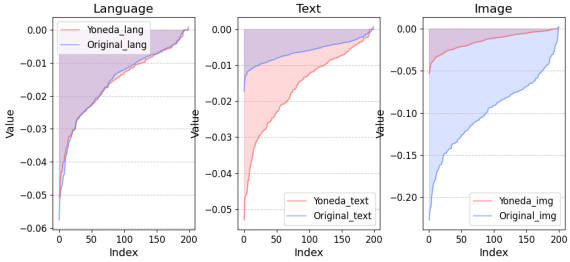


Figure 2: We compute $C(x,y) * C(y,z) - C(x,z)$ across the three categories

4.5. Improvement in CLIP scores

We utilize the CLIP scores to produce a similarity score between the text prompts and the real images. We compute the average CLIP scores of the best K samples and the worst K samples in the test dataset. The best K and the worst K are decided based on (a) the cosine similarities of the samples and (b) loss of a functor that is trained to predict the Jaccard similarity in the text category. When K is set to $1/10^{th}$ of the samples then we find an improvement in the clip scores. However, the overall correlation coefficient is somewhat smaller.

Table 2: Improvement of the CLIP scores with respect to the different metrics

	Metric: Similarity in Yoneda Embeddings	Metric: Test loss on a trained functor
Best-K CLIP scores	0.336	0.334
Worst-K CLIP scores	0.322	0.316
Improvement	4.3%	5.7%

5. Training morphisms while preserving distances

We trained MLPs as morphisms on text and image embeddings to satisfy the reflexivity, transitivity, and neighborhood(prompt IOU3) constraints. We trained a Vision MLP and a language MLP each with two different 3-layer non-commutative branches to map (x,y) or (y,z) to another manifold and compute the cosine similarity of the pair. We feed each text pair and image pair in the combination of each (x, y, z) triplet separately into the two MLPs. For text MLP we flatten the embedding matrix (seq_len, hid_dim). The loss is the MSE with prompt IOU For vision MLP, the loss is the three constraints with different coefficients plus the MSE with the output similarity of the text MLP. We train text MLP for 30 epochs, freeze it and then train the image MLP for 30 epochs.

We find that all loss terms go to zero during training except for **neighborhood loss**. Eventually, whether with or without neighborhood loss(we observe degraded correlation with neighborhood loss), the correlation between vision MLP test loss (suggests how well structure is mapped across modalities) and the CLIP score (text-image matching score) is trivial4. Therefore the experiments failed, and it was surprising that we couldn't reproduce the anchor embedding results in [4]. We deduce that it's because they evaluated different model components trained on the same dataset while we attempted to detect cross-modal latent space communication.

Attempted fixes:

Problems identified: 1. dim too large/dim mismatch 2. Training set (20,000) too small compared with laion 5B; any training attempt will fail Fixes: Take only the end of sentence token from text embedding; (matrix to vector) reduce dim by 77 times Further reduce dim by taking modules exposed to rich data from SD. Project textimage embeddings onto joint multimodal space in the first cross attention layer Results: First stage (prompt IOU) loss is higher. Second stage doesn't converge.

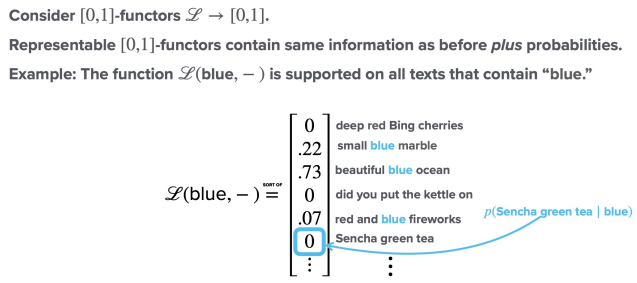


Figure 3: prompt IOU illustration

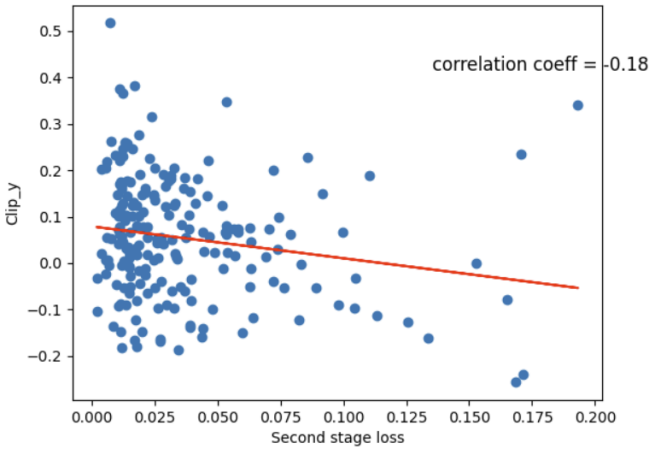


Figure 4: CLIP score correlation with vision MLP loss

6. Open Questions

- Are the CLIP scores reliable?
  No. Even the SOTA open-vocab detection model GLIP is highly unreliable when it comes to detecting composite attributes.
- Perfect correlation with the text and image embeddings, does it mean structure is preserved within the image embeddings?
  No much structure (neighborhood information) is lost from text to image embeddings. Problems could lie in text embedding themselves (Stable Diffusion v2 is only better than v1 in that it has a larger text encoder), so we move to the value/attention matrices derived from text embeddings.
- How can the Yoneda embeddings be used to generate a perfect detector?

References

[1] Tai-Danae Bradley, John Terilla, and Yiannis Vlassopoulos. An enriched category theory of language: from syntax to se-

mantics. *La Matematica*, 1(2):551–580, 2022. 2

[2] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. *arXiv preprint arXiv:2212.05032*, 2022. 1

[3] Bruno Gavranović. Learning functors using gradient descent. *arXiv preprint arXiv:2009.06837*, 2020. 1

[4] Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. Relative representations enable zero-shot latent space communication. *arXiv preprint arXiv:2209.15430*, 2022. 3

[5] Yang Yuan. On the power of foundation models, 2022. 1