Mechanisms of Projective Composition of Diffusion Models

Arwen Bradley [* 1]   Preetum Nakkiran [* 1]   David Berthelot [1]   James Thornton [1]   Joshua M. Susskind [1]

Abstract

We study the theoretical foundations of composition in diffusion models, with a particular focus on out-of-distribution extrapolation and length-generalization. Prior work has shown that composing distributions via linear score combination can achieve promising results, including length-generalization in some cases (Du et al., 2023; Liu et al., 2022). However, our theoretical understanding of how and why such compositions work remains incomplete. In fact, it is not even entirely clear what it means for composition to "work". This paper starts to address these fundamental gaps. We begin by precisely defining one possible desired result of composition, which we call *projective composition*. Then, we investigate: (1) when linear score combinations provably achieve projective composition, (2) whether reverse-diffusion sampling can generate the desired composition, and (3) the conditions under which composition fails. Finally, we connect our theoretical analysis to prior empirical observations where composition has either worked or failed, for reasons that were unclear at the time.
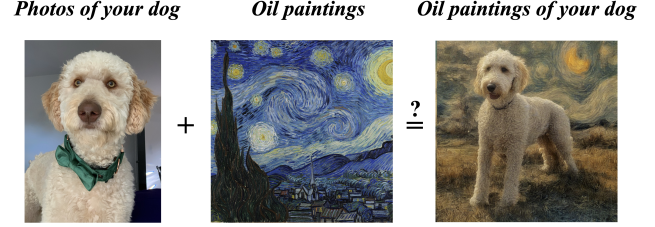
Figure 1: Composing diffusion models via score combination. Given two diffusion models, it is sometimes possible to sample in a way that composes content from one model (e.g. your dog) with style of another model (e.g. oil paintings). We aim to theoretically understand this empirical behavior. Figure generated via score composition with SDXL fine-tuned on the author's dog; details in Appendix C.
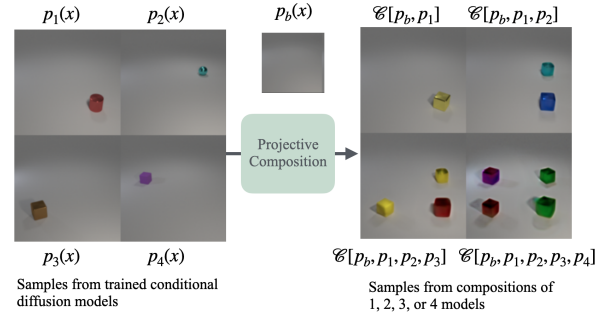

Figure 2: Length-generalization, another capability of composition enabled by our framework. Diffusion models trained to generate a single object conditioned on location (left) can be composed at inference-time to generate images of multiple objects at specified locations (right). Notably, such images are strictly out-of-distribution for the individual models being composed. (Additional samples in Figure 9.)

1. Introduction

The possibility of *composing* different concepts represented by pretrained models has been of both theoretical and practical interest for some time (Jacobs et al., 1991; Hinton, 2002; Du & Kaelbling, 2024), with diverse applications including image and video synthesis (Du et al., 2023; 2020; Liu et al., 2022; 2021; Nie et al., 2021; Yang et al., 2023a; Wang et al., 2024), planning (Ajay et al., 2024; Janner et al., 2022), constraint satisfaction (Yang et al., 2023b), parameter-efficient training (Hu et al., 2021; Ilharco et al., 2022), and many others (Wu et al., 2024; Su et al., 2024; Urain et al., 2023; Anonymous, 2024). One central goal in this field is to build

novel compositions at inference time using only the outputs of pretrained models (either entirely separate models, or different conditionings of a single model), to create generations that are potentially more complex than any model could produce individually. As a concrete example to keep in mind, suppose we have two diffusion models, one trained on your personal photos of your dog and another trained on a collection of oil paintings, and we want to somehow combine these to generate oil paintings of your dog. Note that in

---
*Equal contribution [1]Apple, Cupertino, CA, USA. Correspondence to: Arwen Bradley <arwen_bradley@apple.com>, Preetum Nakkiran <p_nakkiran@apple.com>.

arXiv:2502.04549v1 [cs.LG] 6 Feb 2025

order to achieve this goal, compositions must be able to generate images that are out-of-distribution (OOD) with respect to each of the individual models, since for example, there was no oil painting of your dog in either model's training set. Prior empirical work has shown that this ambitious vision is at least partially achievable in practice. However, the theoretical foundations of how and why composition works in practice, as well as its limitations, are still incomplete.

The goal of this work is to advance our theoretical understanding of composition— we will take a specific family of methods used for composing diffusion models, and we will analyze conditions under which this method provably generates the "correct" composition. Specifically, are there sufficient properties of the distributions we are composing that can guarantee that composition will work "correctly"? And what does correctness even mean, formally?

We focus our study on composing diffusion models by linearly combining their scores, a method introduced by Du et al. (2023); Liu et al. (2022) (though many other interesting constructions are possible, see Section 2). Concretely, suppose we have three separate diffusion models, one for the distribution of dog images p_{dog}, another for oil-paintings p_{oil}, and another unconditional model for generic images p_u. Then, we can use the individual score estimates $\nabla_x \log p(x)$ given by the models to construct a composite score:

$$\nabla_x \log \hat{p}(x) := \tag{1}$$
$$\nabla_x \log p_{dog}(x) + \nabla_x \log p_{oil}(x) - \nabla_x \log p_u(x).$$

This implicitly defines a distribution which we will call a "product composition": $\hat{p}(x) \propto p_{dog}(x)p_{oil}(x)/p_u(x)$. Finally, we can try to sample from $\hat{p}$ by using these scores with a generic score-based sampler, or even reverse-diffusion. This method of composition often achieves good results in practice, yielding e.g. oil paintings of dogs, but it is unclear why it works theoretically.

We are particularly interested in the OOD generalization capabilities of this style of composition. By this we mean the compositional method's ability to generate OOD with respect to each of the individual models being composed – which may be possible even if none of the individual models are themselves capable of OOD generation. A specific desiderata is *length-generalization*, understood as the ability to compose arbitrarily many concepts. For example, consider the CLEVR (Johnson et al., 2017) setting shown in Figure 2. Given conditional models trained on images each containing a single object and conditioned on its location, we want to generate images containing $k > 1$ objects composed in the same scene. How could such length-generalizing composition be possible? Here is one illustrative toy example— consider the following construction, inspired by but slightly different from Du et al. (2023); Liu et al. (2022). Suppose p_b is a distribution of empty background images, and each p_i a

distribution of images with a single object at location i, on an otherwise empty background. Assume all locations we wish to compose are non-overlapping. Then, performing reverse-diffusion sampling using the following score-composition will work — meaning will produce images with k objects at appropriate locations:

$$\nabla_x \log p_b^t(x) + \sum_{i=1}^{k} \underbrace{\left(\nabla_x \log p_i^t(x) - \nabla_x \log p_b^t(x)\right)}_{\text{score delta } \delta_i \, \in \, \mathbb{R}^n}. \tag{2}$$

Above, the notation p_i^t denotes the distribution p_i after time t in the forward diffusion process (see Appendix D). Intuitively this works because during the reverse-diffusion process, the update performed by model i modifies only pixels in the vicinity of location i, and otherwise leaves them identical to the background. Thus the different models do not interact, and the sampler acts as if each model individually "pastes" an object onto an empty background. Formally, sampling works because the score delta vectors δ_i are mutually orthogonal, and in fact have disjoint supports. Notably, we can sample from this composition with a *standard diffusion sampler*, in contrast to Du et al. (2023)'s observations that more sophisticated samplers are necessary. This construction would not be guaranteed to work, however, if the "background" p_b was chosen to be the unconditional distribution p_u (as in Equation 1), a common choice in many prior works (Du et al., 2023; Liu et al., 2022).

The remainder of this paper is devoted to trying to generalize this example as far as possible, and understand both its explanatory power and its limitations. It turns out the core mechanism can be generalized surprisingly far, and does not depend on "orthogonality" as strongly as the above example may suggest. We will encounter some subtle aspects along the way, starting from formal definitions of what it means for composition to succeed — a definition that can capture both composing objects (as in Figure 2), and composing other attributes (such as style + content, in Figure 1).

1.1. Contributions and Organization

In this work we introduce a theoretical framework to help understand the empirical success of certain methods of composing diffusion models, with emphasis on understanding how compositions can sometimes length-generalize. We start by discussing the limitations of several prior definitions of composition in Section 3. In Section 4 we offer a formal definition of "what we want composition to do", given precise information about which aspects we want to compose, which we call *Projective Composition* (Definition 4.1). (Note that there are many other valid notions of composition; we are merely formalizing one particular goal.) Then, we study how projective composition can be achieved. In Section 5 we introduce formal criteria called *Factorized Conditionals* (Definition 5.2), which is a type
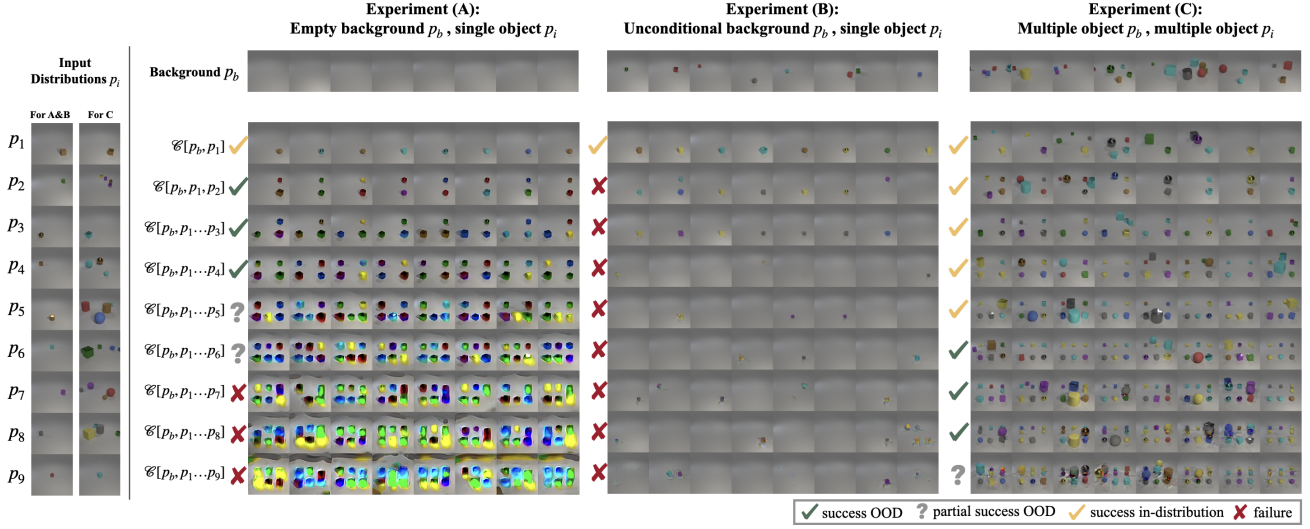
Figure 3: **Attempted compositional length-generalization up to 9 objects.** We attempt to compose via linear score combination the distributions p_1 through p_9 shown on the far left, where each p_i is conditioned on a specific object location as described below. Settings (A) and (C) approximately satisfy the conditions of our theory of projective composition, and thus are expected to length-generalize at least somewhat, while setting (B) does not even approximately satisfy our conditions and indeed fails to length-generalize. **Experiment (A):** In this experiment, the distributions p_i each contain a single object at a fixed location, and the background p_b is empty. In this case any successful composition of more than one object represents length-generalization. We find that composition succeeds up to several objects, but then degrades as number of objects increases (see Section 5.3 for details). **Experiment (B):** Here the distributions p_i are identical to (A), but the background p_b is chosen as the unconditional distribution (i.e. a single object at a random location)— this the "Bayes composition" (Section 3). This composition entirely fails— remarkably, trying to compose many objects often produces no objects! **Experiment (C):** Here each distribution p_i contains an object at a fixed location i, and $0 - 4$ other objects (sampled uniformly) in random locations; see samples at far left. The background distribution p_b is a distribution of $1 - 5$ objects (sampled uniformly) in random locations. In this case length-generalization means composition of more than 5 objects. This composition can length-generalize, but artifacts appear for large numbers of objects. See Section 5.3 for a full discussion.

of independence criteria along both distributions and coordinates. We prove that when this criteria holds, projective composition can be achieved by linearly combining scores (as in Equation 2), and can be sampled via standard reverse-diffusion samplers. In Section 6 we show that parts of this result can be extended much further to apply even in nonlinear feature spaces; but interestingly, even when projective composition is achievable, it may be difficult to sample. We find that in many important cases existing constructions approximately satisfy our conditions, but the theory also helps characterize and explain certain limitations. Finally in Section 8 we discuss how our results can help explain existing experimental results in the literature where composition worked or failed, for reasons that were unclear at the time.

2. Related Work

Single vs. Multiple Model Composition. First, we distinguish the kind of composition we study in this paper from approaches that rely a single model but with OOD conditioning—

ers; for example, passing OOD text prompts to text-to-image models (Nichol et al., 2021; Podell et al., 2023), or works like Okawa et al. (2024); Park et al. (2024). In contrast, we study compositions which recombine the outputs of multiple separate models at inference time, where each model only sees in-distribution conditionings. Among compositions involving multiple models, many different variants have been explored. Some are inspired by logical operators like AND and OR, which are typically implemented as product $p_0(x)p_1(x)$ and sum $p_0(x)+p_1(x)$ (Du et al., 2023; Du & Kaelbling, 2024; Liu et al., 2022). Some composition methods are based on diffusion models, while others use energy-based models (Du et al., 2020; 2023; Liu et al., 2021) or densities (Skreta et al., 2024). In this work, we focus specifically on product-style compositions implemented with diffusion models via a linear combinations of scores as in Du et al. (2023); Liu et al. (2022). Our goal is not to propose a new method of composition but to improve theoretical understanding of existing methods.

Learning and Generalization. In this work we focus only on mathematical aspects of composition, and we do not consider any learning-theoretic aspects such as inductive bias or sample complexity. Our work is thus complementary to Kamb & Ganguli (2024), which studies how a type of compositional generalization can arise from inductive bias in the learning procedure. Additional related works are discussed in Appendix A.

3. Prior Definitions of Composition

In this section we will describe why two popular mathematical definitions of composition are insufficient for our purposes: the "simple product" definition, and the Bayes composition. Specifically, neither of these notions can describe the outcome of the CLEVR length-generalization experiment from Figure 2. Our observations here will thus motivate us to propose a new definition of composition, in the following section. As a running example, we will consider a subset of the CLEVR experiment from Figure 2. Suppose we are trying to compose two distributions p_1, p_2 of images each containing a single object in an otherwise empty scene, where the object is in the lower-left corner under p_1, and the upper-right corner under p_2. We would like the composed distribution $\hat{p}$ to place objects in at least the lower-left and upper-right, simultaneously.

3.1. The Simple Product

The simple product is perhaps the most familiar type of composition: Given two distributions p_1 and p_2 over $\mathbb{R}^n$, the simple product is defined[1] as $\hat{p}(x) \propto p_1(x)p_2(x)$. The simple product can represent some interesting types of composition, but it has a key limitation: the composed distribution can never be truly "out-of-distribution" w.r.t. p_1 or p_2, since $\hat{p}(x) = 0$ whenever $p_1(x) = 0$ or $p_2(x) = 0$. This presents a problem for our CLEVR experiment. Using the simple product definition, we must have $\hat{p}(x) = 0$ for any image x with two objects, since neither p_1 nor p_2 was supported on images with two objects. Therefore, the simple product definition cannot represent our desired composition.

3.2. The Bayes Composition

Another candidate definition for composition, which we will call the "Bayes composition", was introduced and studied by Du et al. (2023); Liu et al. (2022). The Bayes composition is theoretically justified when the desired composed distribution is formally a conditional distribution of the model's training distribution. However, it is not formally capable of generating truly out-of-distribution samples, as our example below will illustrate.

Let us attempt to apply the Bayes composition methodology to our CLEVR example. We interpret our two distributions p_1, p_2 as conditional distributions, conditioned on an object appearing in the lower-left or upper-right, respectively. Thus we write $p(x|c_1) \equiv p_1(x)$, where c_1 is the event that an object appears in the lower-left of image x, and c_2 the event an object appears in the upper-right. Now, since we want both objects simultaneously, we define the composition as $\hat{p}(x) := p(x|c_1, c_2)$. Because the two events c_1 and c_2 are conditionally independent given x (since they are deterministic functions of x), we can compute $\hat{p}$ in terms of the individual conditionals:

$$\hat{p}(x) := p(x|c_1, c_2) \propto p(x|c_1)p(x|c_2)/p(x). \quad (3)$$

Equivalently in terms of scores: $\nabla_x \log \hat{p}_t(x) := \nabla_x \log p(x|c_1) + \nabla_x \log p(x|c_2) - \nabla_x \log p(x)$. Line (3) thus serves as our definition of the Bayes composition $\hat{p}$, in terms of the conditional distributions $p(x|c_1)$ and $p(x|c_2)$, and the unconditional $p(x)$.

The definition of composition above seems natural: we want both objects to appear simultaneously, so let us simply condition on both these events. However, there is an obvious error in the conclusion: $\hat{p}(x)$ must be 0 whenever $p(x|c_1)$ or $p(x|c_2)$ is zero (by Line 3). Since neither conditional distribution have support on images with two objects, the composition $\hat{p}$ cannot contain images of two objects either.

Where did this go wrong? The issue is: $p(x|c_1, c_2)$ is not well-defined in our case. We intuitively imagine some unconditional distribution $p(x)$ which allows both objects simultaneously, but no such distribution has been defined, or encountered by the models during training. Thus, the definition of $\hat{p}$ in Line (3) does not actually correspond to our intuitive notion of "conditioning on both objects at once." More generally, this example illustrates how the Bayes composition cannot produce truly out-of-distribution samples, with respect to the distributions being composed.[2] Figure 3b shows that the Bayes composition does not always work experimentally either: for diffusion models trained in a CLEVR setting similar to Figure 2, the Bayes composition of $k > 1$ locations typically fails to produce k objects (and sometimes produces zero). The difficulties discussed lead us to propose a precise definition of what we actually "want" composition to do in this case.

4. Our Proposal: Projective-Composition

We now present our formal definition of what it means to "correctly compose" distributions. Our main insight here is, a realistic definition of composition should not purely be a function of distributions $\{p_1, p_2, \dots\}$, in the way the simple

---

[1]The geometric mean $\sqrt{p_1(x)p_2(x)}$ is also often used; our discussion applies equally to this as well.

[2]Although Du et al. (2023) use the Bayes composition to achieve a kind of length-generalization, our discussion shows that the Bayes justification does not explain the experimental results.
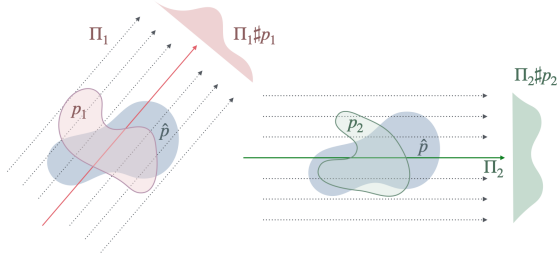
Figure 4: Distribution $\hat{p}$ is a projective composition of p_1 and p_2 w.r.t. projection functions (Π_1, Π_2), because $\hat{p}$ has the same marginals as p_1 when both are post-processed by Π_1, and analogously for p_2.
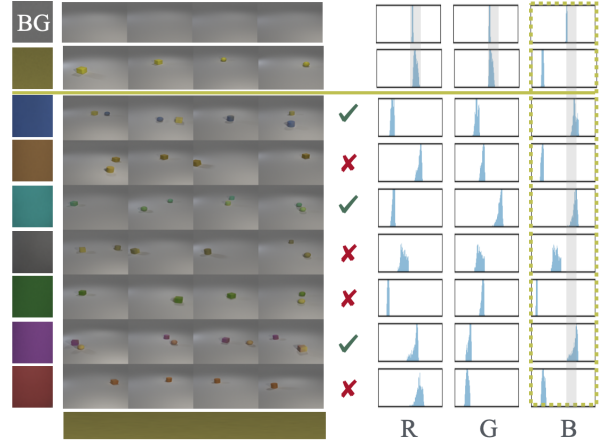

Figure 5: **Composing yellow objects with objects of other colors.** Yellow objects successfully compose with blue, cyan and magenta objects but not with brown, gray, green, or red objects. Per the histograms (left), in RGB-colorspace yellow has R, G distributed like the background (gray) while B has a distinct distribution peaked closer to zero. Taking $M_{\text{yellow}} \approx \{B\}$, Theorem 5.3 predicts that standard diffusion can sample from compositions of yellow with any color where the B channel is distributed like the background: namely, blue, cyan, magenta per the histograms. (Other colors may theoretically compose per Theorem 6.1, but be difficult to sample.) (Additional samples in Figure 10.)

product $\hat{p}(x) = p_1(x)p_2(x)$ is purely a function of p_1, p_2. We must also somehow specify *which aspects* of each distribution we care about preserving in the composition. For example, informally, we may want a composition that mimics the style of p_1 and the content of p_2. Our definition below of *projective composition* allows us this flexibility.

Roughly speaking, our definition requires specifying a "feature extractor" $\Pi_i : \mathbb{R}^n \to \mathbb{R}_k$ associated with every distribution p_i. These functions can be arbitrary, but we usually imagine them as projections[3] in some feature-space, e.g, $\Pi_1(x)$ may be a transform of x which extracts only its style, and $\Pi_2(x)$ a transform which extracts only its content. Then, a projective composition is any distribution $\hat{p}$ which "looks like" distribution p_i when both are viewed through Π_i (see Figure 4). Formally:

Definition 4.1 (Projective Composition). *Given a collection of distributions $\{p_i\}$ along with associated "projection" functions $\{\Pi_i : \mathbb{R}^n \to \mathbb{R}^k\}$, we call a distribution $\hat{p}$ a* projective composition *if[4]*

$$\forall i : \quad \Pi_i \sharp \hat{p} = \Pi_i \sharp p_i. \tag{4}$$

That is, when $\hat{p}$ is projected by each Π_i, it yields marginals identical to those of p_i.

There are a few aspects of this definition worth emphasizing, which are conceptually different from many prior notions of composition. First, our definition above does not *construct* a composed distribution; it merely specifies what properties the composition must have. For a given set of $\{(p_i, \Pi_i)\}$, there may be many possible distributions $\hat{p}$ which are projective compositions; or in other cases, a projective composition may not even exist. Separately, the definition of projective composition does not posit any sort of "true" underlying distribution, nor does it require that

the distributions p_i are conditionals of an underlying joint distribution. In particular, projective compositions can be truly "out of distribution" with respect to the p_i: $\hat{p}$ can be supported on samples x where none of the p_i are supported.

Examples. We have already discussed the style+content composition of Figure 1 as an instance of projective composition. Another even simpler example to keep in mind is the following coordinate-projection case. Suppose we take $\Pi_i : \mathbb{R}^n \to \mathbb{R}$ to be the projection onto the i-th coordinate. Then, a projective composition of distributions $\{p_i\}$ with these associated functions $\{\Pi_i\}$ means: a distribution where the first coordinate is marginally distributed identically to the first coordinate of p_1, the second coordinate is marginally distributed as p_2, and so on. (Note, we do not require any independence between coordinates). This notion of composition would be meaningful if, for example, we are already working in some disentangled feature space, where the first coordinate controls the style of the image the second coordinate controls the texture, and so on. The CLEVR length-generalization example from Figure 2 can also be described as a projective composition in almost an identical way, by letting $\Pi_i : \mathbb{R}^n \to \mathbb{R}^k$ be a restriction onto the set of pixels neighboring location i. We describe this

---

[3]We use the term "projection" informally here, to convey intuition; these functions Π_i are not necessarily coordinate projections, although this is an important special case (Section 5).

[4]The notation $\sharp$ refers to push-forward of a probability measure.

explicitly later in Section 5.3.

5. Simple Construction of Projective Compositions

It is not clear apriori that projective compositional distributions satisfying Definition 4.1 ever exist, much less that there is any straightforward way to sample from them. To explore this, we first restrict attention to perhaps the simplest setting, where the projection functions $\{\Pi_i\}$ are just coordinate restrictions. This setting is meant to generalize the intuition we had in the CLEVR example of Figure 2, where different objects were composed in disjoint regions of the image. We first define the construction of the composed distribution, and then establish its theoretical properties.

5.1. Defining the Construction

Formally, suppose we have a set of distributions $(p_1, p_2, \ldots, p_k)$ that we wish to compose; in our running CLEVR example, each p_i is the distribution of images with a single object at position i. Suppose also we have some reference distribution p_b, which can be arbitrary, but should be thought of as a "common background" to the p_is. Then, one popular way to construct a composed distribution is via the *compositional operator* defined below. (A special case of this construction is used in Du et al. (2023), for example.)

Definition 5.1 (Composition Operator). *Define the composition operator $\mathcal{C}$ acting on an arbitrary set of distributions $(p_b, p_1, p_2, \ldots)$ by*

$$\mathcal{C}[\vec{p}] := \mathcal{C}[p_b, p_1, p_2, \ldots](x) := \frac{1}{Z} p_b(x) \prod_i \frac{p_i(x)}{p_b(x)}, \quad (5)$$

where Z is the appropriate normalization constant. We name $\mathcal{C}[\vec{p}]$ the composed distribution*, and the score of $\mathcal{C}[\vec{p}]$ the* compositional score*:*

$$\nabla_x \log \mathcal{C}[\vec{p}](x) \qquad (6)$$
$$= \nabla_x \log p_b(x) + \sum_i \left(\nabla_x \log p_i(x) - \nabla_x \log p_b(x)\right).$$

Notice that if p_b is taken to be the unconditional distribution then this is exactly the Bayes-composition.

5.2. When does the Composition Operator Work?

We can always apply the composition operator to any set of distributions, but when does this actually yield a "correct" composition (according to Definition 4.1)? One special case is when each distribution p_i is "active" on a different, non-overlapping set of coordinates. We formalize this property below as *Factorized Conditionals* (Definition 5.2). The idea is, each distribution p_i must have a particular set of "mask" coordinates $M_i \subseteq [n]$ which it samples in a characteristic

way, while independently sampling all other coordinates from a common background distribution. If a set of distributions $(p_b, p_1, p_2, \ldots)$ has this *Factorized Conditional* structure, then the composition operator will produce a projective composition (as we will prove below).

Definition 5.2 (Factorized-Conditionals). *We say a set of distributions $(p_b, p_1, p_2, \ldots p_k)$ over $\mathbb{R}^n$ are* Factorized Conditionals *if there exists a partition of coordinates $[n]$ into disjoint subsets $M_b, M_1, \ldots M_k$ such that:*

1. *$(x|_{M_i}, x|_{M_i^c})$ are independent under p_i.*

2. *$(x|_{M_b}, x|_{M_1}, x|_{M_2}, \ldots, x|_{M_k})$ are mutually independent under p_b.*

3. *$p_i(x|_{M_i^c}) = p_b(x|_{M_i^c})$.*

Equivalently, if we have:

$$p_i(x) = p_i(x|_{M_i})p_b(x|_{M_i^c}), \text{ and} \qquad (7)$$
$$p_b(x) = p_b(x|_{M_b}) \prod_{i \in [k]} p_b(x|_{M_i}).$$

Equation (7) means that each p_i can be sampled by first sampling $x \sim p_b$, and then overwriting the coordinates of M_i according to some other distribution (which can be specific to distribution i). For instance, the experiment of Figure 2 intuitively satisfies this property, since each of the conditional distributions could essentially be sampled by first sampling an empty background image (p_b), then "pasting" a random object in the appropriate location (corresponding to pixels M_i). If a set of distributions obey this Factorized Conditional structure, then we can prove that the composition operator $\mathcal{C}$ yields a correct projective composition, and reverse-diffusion correctly samples from it. Below, let N_t denote the noise operator of the diffusion process[5] at time t.

Theorem 5.3 (Correctness of Composition). *Suppose a set of distributions $(p_b, p_1, p_2, \ldots p_k)$ satisfy Definition 5.2, with corresponding masks $\{M_i\}_i$. Consider running the reverse-diffusion SDE using the following compositional scores at each time t:*

$$s_t(x_t) := \nabla_x \log \mathcal{C}[p_b^t, p_1^t, p_2^t, \ldots](x_t), \qquad (8)$$

where $p_i^t := N_t[p_i]$ are the noisy distributions. Then, the distribution of the generated sample x_0 at time $t = 0$ is:

$$\hat{p}(x) := p_b(x|_{M_b}) \prod_i p_i(x|_{M_i}). \qquad (9)$$

In particular, $\hat{p}(x|_{M_i}) = p_i(x|_{M_i})$ for all i, and so $\hat{p}$ is a projective composition with respect to projections $\{\Pi_i(x) := x|_{M_i}\}_i$, per Definition 4.1.

---

[5] Our results are agnostic to the specific diffusion noise-schedule and scaling used.

Unpacking this, Line 9 says that the final generated distribution $\hat{p}(x)$ can be sampled by first sampling the coordinates M_b according to p_b (marginally), then independently sampling coordinates M_i according to p_i (marginally) for each i. Similarly, by assumption, $p_i(x)$ can be sampled by first sampling the coordinates M_i in some specific way, and then independently sampling the remaining coordinates according to p_b. Therefore Theorem 5.3 says that $\hat{p}(x)$ samples the coordinates M_i *exactly as they would be sampled* by p_i, for each i we wish to compose.

Proof. (Sketch) Since $\vec{p}$ satisfies Definition 5.2, we have

$$\mathcal{C}[\vec{p}](x) := p_b(x) \prod_i \frac{p_i(x)}{p_b(x)} = p_b(x) \prod_i \frac{p_b(x_t|_{M_i^c})p_i(x|_{M_i})}{p_b(x|_{M_i^c})p_b(x|_{M_i})}$$

$$= p_b(x) \prod_i \frac{p_i(x|_{M_i})}{p_b(x|_{M_i})} = p_b(x|_{M_b}) \prod_i p_i(x_t|_{M_i}) := \hat{p}(x).$$

The sampling guarantee follows from the commutativity of composition with the diffusion noising process, i.e. $\mathcal{C}[\vec{p^t}] = N_t[\mathcal{C}[\vec{p}]]$. The complete proof is in Appendix G. □

Remark 5.4. In fact, Theorem 5.3 still holds under any orthogonal transformation of the variables, because the diffusion noise process commutes with orthogonal transforms. We formalize this as Lemma 7.1.

Remark 5.5. Compositionality is often thought of in terms of orthogonality between scores. Definition 5.2 implies orthogonality between the score differences that appear in the composed score (6): $\nabla_x \log p_i^t(x_t) - \nabla_x \log p_b^t(x_t)$, but the former condition is strictly stronger (c.f. Appendix F).

Remark 5.6. Notice that the composition operator $\mathcal{C}$ can be applied to a set of Factorized Conditional distributions without knowing the coordinate partition $\{M_i\}$. That is, we can compose distributions and compute scores without knowing apriori exactly "how" these distributions are supposed to compose (i.e. which coordinates p_i is active on). This is already somewhat remarkable, and we will see a much stronger version of this property in the next section.

Importance of background. Our derivations highlight the crucial role of the background distribution p_b for the composition operator (Definition 5.1). While prior works have taken p_b to be an unconditional distribution and the p_i's its associated conditionals, our results suggest this is not always the optimal choice – in particular, it may not satisfy a Factorized Conditional structure (Definition 5.2). Figure 3 demonstrates this empirically: settings (a) and (b) attempt to compose the same distributions using different backgrounds – empty (a) or unconditional (b) – with very different results.

5.3. Approximate Factorized Conditionals in CLEVR.

In Figure 3 we explore compositional length-generalization (or lack thereof) in three different setting, two of which

(Figure 3a and 3c) approximately satisfy Definition 5.2. In this section we explicitly describe how our definition of Factorized Conditionals approximately captures the CLEVR settings of Figures 3a and 3c. The setting of 3b does not satisfy our conditions, as discussed in Section 3.

Single object distributions with empty background. This is the setting of both Figure 2 and Figure 3a. The background distribution p_b over n pixels is images of an empty scene with no objects. For each $i \in \{1, \ldots, L\}$ (where $L = 4$ in Figure 2 and $L = 9$ in Figure 3a), define the set $M_i \subset [n]$ as the set of pixel indices surrounding location i. (M_i should be thought of as a "mask" that that masks out objects at location i). Let $M_b := (\cup_i M_i)^c$ be the remaining pixels in the image. Then, we claim the distributions $(p_b, p_1, \ldots, p_L)$ form approximately Factorized Conditionals, with corresponding coordinate partition $\{M_i\}$. This is essentially because each distribution p_i matches the background p_b on all pixels except those surrounding location i (further detail in Appendix B.2). Note, however, that the conditions of Definition 5.2 do not *exactly* hold in the experiment of Figure 2 – there is still some dependence between the masks M_i, since objects can cast shadows or even occlude each other. Empirically, these deviations have greater impact when composing many objects, as seen in Figure 3a.

Bayes composition with cluttered distributions. In Figure 3c we replicate CLEVR experiments in Du et al. (2023); Liu et al. (2022) where the images contain many objects (1-5) and the conditions label the location of one randomly-chosen object. It turns out the unconditional together with the conditionals can approximately act as Factorized Conditionals in "cluttered" settings like this one. The intuition is that if the conditional distributions each contain one specific object plus many independently sampled random objects ("clutter"), then the unconditional distribution *almost* looks like independently sampled random objects, which together with the conditionals *would* satisfy Definition 5.2 (further discussion in Appendix B.2 and E). This helps to explain the length-generalization observed in Liu et al. (2022) and verified in our experiments (Figure 3c).

6. Projective Composition in Feature Space

So far we have focused on the setting where the projection functions Π_i are simply projections onto coordinate subsets M_i in the native space (e.g. pixel space). This covers simple examples like Figure 2 but does not include more realistic situations such as Figure 1, where the properties to be composed are more abstract. For example a property like "oil painting" does not correspond to projection onto a specific subset of pixels in an image. However, we may hope that there exists some conceptual feature space in which "oil painting" does correspond to a particular subset of variables. In this section, we extend our results to the case where
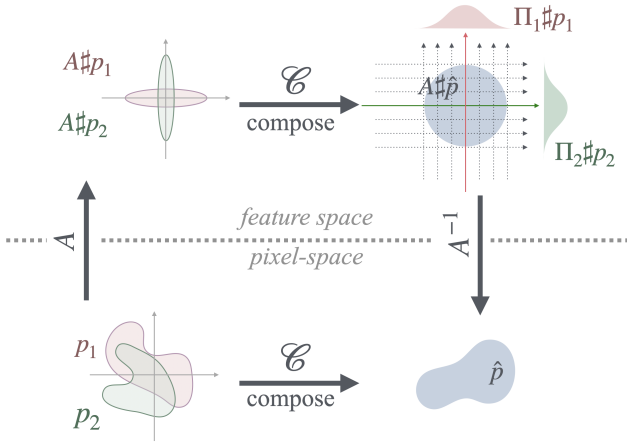
Figure 6: A commutative diagram illustrating Theorem 6.1. Performing composition in pixel space is equivalent to encoding into a feature space ($\mathcal{A}$), composing there, and decoding back to pixel space ($\mathcal{A}^{-1}$).

the composition occurs in some conceptual feature space, and each distribution to be composed corresponds to some particular subset of *features*.

Our main result is a featurized analogue of Theorem 5.3: if there exists *any* invertible transform $\mathcal{A}$ mapping into a feature space where Definition 5.2 holds, then the composition operator (Definition 5.1) yields a projective composition in this feature space, as shown in Figure 6.

Theorem 6.1 (Feature-space Composition). *Given distributions $\vec{p} := (p_b, p_1, p_2, \ldots p_k)$, suppose there exists a diffeomorphism $\mathcal{A} : \mathbb{R}^n \to \mathbb{R}^n$ such that $(\mathcal{A}\sharp p_b, \mathcal{A}\sharp p_1, \ldots \mathcal{A}\sharp p_k)$ satisfy Definition 5.2, with corresponding partition $M_i \subseteq [n]$. Then, the composition $\hat{p} := \mathcal{C}[\vec{p}]$ satisfies:*

$$\mathcal{A}\sharp\hat{p}(z) \equiv (\mathcal{A}\sharp p_b(z))|_{M_b} \prod_{i=1}^{k} (\mathcal{A}\sharp p_i(z))|_{M_i}. \quad (10)$$

Therefore, $\hat{p}$ is a projective composition of $\vec{p}$ w.r.t. projection functions $\{\Pi_i(x) := \mathcal{A}(x)|_{M_i}\}$.

This theorem is remarkable because it means we can compose distributions $(p_b, p_1, p_2, \ldots)$ in the base space, and this composition will "work correctly" in the feature space automatically (Equation 10), without us ever needing to compute or even know the feature transform $\mathcal{A}$ explicitly.

Theorem 6.1 may apriori seem too strong to be true, since it somehow holds for all feature spaces $\mathcal{A}$ simultaneously. The key observation underlying Theorem 6.1 is that the composition operator $\mathcal{C}$ behaves well under reparameterization.

Lemma 6.2 (Reparameterization Equivariance). *The composition operator of Definition 5.1 is reparameterization-equivariant. That is, for all diffeomorphisms $\mathcal{A} : \mathbb{R}^n \to \mathbb{R}^n$*

and all tuples of distributions $\vec{p} = (p_b, p_1, p_2, \ldots, p_k)$,

$$\mathcal{C}[\mathcal{A}\sharp\vec{p}] = \mathcal{A}\sharp\mathcal{C}[\vec{p}]. \quad (11)$$

This lemma is potentially of independent interest: equivariance distinguishes the composition operator from many other common operators (e.g. the simple product). Lemma 6.2 and Theorem 6.1 are proved in Appendix H.

7. Sampling from Compositions.

The feature-space Theorem 6.1 is weaker than Theorem 5.3 in one important way: it does not provide a sampling algorithm. That is, Theorem 6.1 guarantees that $\hat{p} := \mathcal{C}[\vec{p}]$ is a projective composition, but does not guarantee that reverse-diffusion is a valid sampling method.

There is one special case where diffusion sampling *is* guaranteed to work, namely, for orthogonal transforms (which can seen as a straightforward extension of the coordinate-aligned case of Theorem 5.3):

Lemma 7.1 (Orthogonal transform enables diffusion sampling). *If the assumptions of Lemma 6.1 hold for $\mathcal{A}(x) = Ax$, where A is an orthogonal matrix, then running a reverse diffusion sampler with scores $s_t = \nabla_x \log \mathcal{C}[\vec{p}^t]$ generates the composed distribution $\hat{p} = \mathcal{C}[\vec{p}]$ satisfying (10).*

The proof is given in Appendix I.

However, for general invertible transforms, we have no such sampling guarantees. Part of this is inherent: in the feature-space setting, the diffusion noise operator N_t no longer commutes with the composition operator $\mathcal{C}$ in general, so scores of the noisy composed distribution $N_t[\mathcal{C}[\vec{p}]]$ cannot be computed from scores of the noisy base distributions $N_t[\vec{p}]$. Nevertheless, one may hope to sample from the distribution $\hat{p}$ using other samplers besides diffusion, such as annealed Langevin Dynamics or Predictor-Corrector methods (Song et al., 2020). We find that the situation is surprisingly subtle: composition $\mathcal{C}$ produces distributions which are in some cases easy to sample (e.g. with diffusion), yet in other cases apparently hard to sample. For example, in the setting of Figure 5, our Theorem 6.1 implies that all pairs of colors should compose equally well at time $t = 0$, since there exist diffeomorphisms (indeed, linear transforms) between different colors. However, as we saw, the diffusion sampler fails to sample from compositions of non-orthogonal colors— and empirically, even more sophisticated samplers such as Predictor-Correctors also fail in this setting. At first glance, it may seem odd that composed distributions are so hard to sample, when their constituent distributions are relatively easy to sample. One possible reason for this below is that the composition operator has extremely poor Lipchitz constant, so it is possible for a set of distributions $\vec{p}$ to "vary smoothly" (e.g. diffusing over time) while their

"photo of a dog" + "photo of a horse"   *"photo of a dog" + "photo, with red hat"*

Figure 7: **Composing Entangled Concepts.** The left image composes the text-conditions "photo of a dog" with "photo of a horse", which both control the subject of the image, and produces unexpected results. In contrast, the right image composes "photo of a dog" with "photo, with red hat," which intuitively correspond to disentangled features. Both samples from SDXL using score-composition with an unconditional background; details in Appendix C.

composition $\mathcal{C}[\vec{p}]$ changes abruptly. We formalize this in Lemma 7.2 (further discussion and proof in Appendix J).

Lemma 7.2 (Composition Non-Smoothness)**.** *For any set of distributions* $\{p_b, p_1, p_2, \ldots, p_k\}$*, and any noise scale* $t := \sigma$*, define the noisy distributions* $p_i^t := N_t[p_i]$*, and let* q^t *denote the composed distribution at time* t*:* $q^t := \mathcal{C}[\vec{p}^t]$*. Then, for any choice of* $\tau > 0$*, there exist distributions* $\{p_b, p_1, \ldots p_k\}$ *over* $\mathbb{R}^n$ *such that*

1. *For all* i*, the annealing path of* p_i *is* $\mathcal{O}(1)$*-Lipshitz:* $\forall t, t' : W_2(p_i^t, p_i^{t'}) \leq \mathcal{O}(1)|t - t'|$*.*

2. *The annealing path of* q *has Lipshitz constant at least* $\Omega(\tau^{-1})$*:* $\exists t, t' : W_2(q^t, q^{t'}) \geq \frac{|t-t'|}{2\tau}$*.*

8. Connections with Prior Observations

We have presented a mathematical theory of composition. Although this theoretical model is a simplification of reality (we do not claim its assumptions hold exactly in practice) we believe the spirit of our results carries over to practical settings, and can help understand empirical observations from prior work. We now discuss some of these connections.

Independence Assumptions and Disentangled Features. Our theory relies on a type of independence between distributions, related to orthogonality between scores, which we formalize as Factorized Conditionals. While such conditional structure typically does not exist in pixel-space, it is plausible that a factorized structure exists in an appropriate *feature space*, as permitted by our theory (Section 6). In particular, a feature space and distribution with perfectly "disentangled" features would satisfy our assumptions. Conversely, if distributions are not appropriately disentangled,

our theory predicts that linear score combinations will fail to compose correctly. This effect is well-known; see Figure 7 for an example; similar failure cases are highlighted in Liu et al. (2022) as well (such as "A bird" failing to compose with "A flower"). Regarding successful cases, style and content compositions consistently work well in practice, and are often taken to be disentangled features (e.g. Karras (2019); Gatys et al. (2016); Zhu et al. (2017)). Finally, similar in spirit to our theory, methods for successful composition in practice such as LoRA task arithmetic (Zhang et al., 2023a; Ilharco et al., 2022), typically require some type of approximate "concept-space orthogonality".

Text conditioning with location information. Conditioning on location is a straightforward way to achieve factorized conditionals (provided the objects in different locations are approximately independent) since the required disjointness already holds in pixel-space. Many successful text-to-image compositions in Liu et al. (2022) use location information in the prompts, either explicitly (e.g. "A blue bird on a tree" + "A red car behind the tree") or implicitly ("A horse" + "A yellow flower field"; since horses are typically in the foreground and fields in the background).

Unconditional Backgrounds. Most prior works on diffusion composition use the Bayes composition, with substantial practical success. As discussed in Section 5.3, Bayes composition may be approximately projective in "cluttered" settings, helping to explain its practical success in text-to-image settings, where images often contain many different possible objects and concepts.

9. Conclusion

In this work, we have developed a theory of one possible mechanism of composition in diffusion models. We study how composition can be defined, and sufficient conditions for it to be achieved. Our theory can help understand a range of diverse compositional phenomena in both synthetic and practical settings, and we hope it will inspire further work on foundations of composition.

Acknowledgements

References

Ajay, A., Han, S., Du, Y., Li, S., Gupta, A., Jaakkola, T., Tenenbaum, J., Kaelbling, L., Srivastava, A., and Agrawal, P. Compositional foundation models for hierarchical planning. *Advances in Neural Information*

Processing Systems, 36, 2024.

Anonymous. Scaling in-the-wild training for diffusion-based illumination harmonization and editing by imposing consistent light transport. In *Submitted to The Thirteenth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=u1cQYxRI1H. under review.

Delon, J. and Desolneux, A. A wasserstein-type distance in the space of gaussian mixture models. *SIAM Journal on Imaging Sciences*, 13(2):936–970, 2020.

Du, Y. and Kaelbling, L. Compositional generative modeling: A single model is not all you need. *arXiv preprint arXiv:2402.01103*, 2024.

Du, Y., Li, S., and Mordatch, I. Compositional visual generation and inference with energy based models. *arXiv preprint arXiv:2004.06030*, 2020.

Du, Y., Durkan, C., Strudel, R., Tenenbaum, J. B., Dieleman, S., Fergus, R., Sohl-Dickstein, J., Doucet, A., and Grathwohl, W. S. Reduce, reuse, recycle: Compositional generation with energy-based diffusion models and mcmc. In *International conference on machine learning*, pp. 8489–8510. PMLR, 2023.

Gatys, L. A., Ecker, A. S., and Bethge, M. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2414–2423, 2016.

Gregor, K., Danihelka, I., Graves, A., Rezende, D. J., and Wierstra, D. Draw: A recurrent neural network for image generation, 2015. URL https://arxiv.org/abs/1502.04623.

Hinton, G. E. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002.

Ho, J. and Salimans, T. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Ilharco, G., Ribeiro, M. T., Wortsman, M., Gururangan, S., Schmidt, L., Hajishirzi, H., and Farhadi, A. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*, 2022.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

Janner, M., Du, Y., Tenenbaum, J. B., and Levine, S. Planning with diffusion for flexible behavior synthesis. *arXiv preprint arXiv:2205.09991*, 2022.

Johnson, J., Hariharan, B., Van Der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., and Girshick, R. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2901–2910, 2017.

Kamb, M. and Ganguli, S. An analytic theory of creativity in convolutional diffusion models. *arXiv preprint arXiv:2412.20292*, 2024.

Karras, T. A style-based generator architecture for generative adversarial networks. *arXiv preprint arXiv:1812.04948*, 2019.

Karras, T., Aittala, M., Lehtinen, J., Hellsten, J., Aila, T., and Laine, S. Analyzing and improving the training dynamics of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24174–24184, 2024.

Liu, N., Li, S., Du, Y., Tenenbaum, J., and Torralba, A. Learning to compose visual relations. *Advances in Neural Information Processing Systems*, 34:23166–23178, 2021.

Liu, N., Li, S., Du, Y., Torralba, A., and Tenenbaum, J. B. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision*, pp. 423–439. Springer, 2022.

Liu, Y., Zhang, Y., Jaakkola, T., and Chang, S. Correcting diffusion generation through resampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8713–8723, 2024.

McAllister, D., Tancik, M., Song, J., and Kanazawa, A. Decentralized diffusion models. *arXiv preprint arXiv:2501.05450*, 2025.

Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., and Chen, M. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

Nie, W., Vahdat, A., and Anandkumar, A. Controllable and compositional generation with latent-space energy-based models. *Advances in Neural Information Processing Systems*, 34:13497–13510, 2021.

Okawa, M., Lubana, E. S., Dick, R., and Tanaka, H. Compositional abilities emerge multiplicatively: Exploring diffusion models on a synthetic task. *Advances in Neural Information Processing Systems*, 36, 2024.

Parisi, G. Correlation functions and computer simulations. *Nuclear Physics B*, 180(3):378–384, 1981.

Park, C. F., Okawa, M., Lee, A., Lubana, E. S., and Tanaka, H. Emergence of hidden capabilities: Exploring learning dynamics in concept space. *arXiv preprint arXiv:2406.19370*, 2024.

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., and Rombach, R. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Robert, C. P., Casella, G., and Casella, G. *Monte Carlo statistical methods*, volume 2. Springer, 1999.

Roberts, G. O. and Tweedie, R. L. Exponential convergence of langevin distributions and their discrete approximations, 1996.

Rossky, P. J., Doll, J. D., and Friedman, H. L. Brownian dynamics as smart monte carlo simulation. *The Journal of Chemical Physics*, 69(10):4628–4633, 1978.

Skreta, M., Atanackovic, L., Bose, A. J., Tong, A., and Neklyudov, K. The superposition of diffusion models using the it\ˆ o density estimator. *arXiv preprint arXiv:2412.17762*, 2024.

Sohl-Dickstein, J., Weiss, E. A., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. *CoRR*, abs/1503.03585, 2015. URL http://arxiv.org/abs/1503.03585.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=St1giarCHLP.

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. URL https://arxiv.org/pdf/2011.13456.pdf.

Stracke, N., Baumann, S. A., Susskind, J., Martin, M. A. B., and Ommer, B. Ctrloralter: Conditional loradapter for efficient zero-shot control & altering of t2i models. In *ECCV*, 2024. URL https://arxiv.org/abs/2405.07913.

Su, J., Liu, N., Wang, Y., Tenenbaum, J. B., and Du, Y. Compositional image decomposition with diffusion models. *arXiv preprint arXiv:2406.19298*, 2024.

Urain, J., Li, A., Liu, P., D'Eramo, C., and Peters, J. Composable energy policies for reactive motion generation and reinforcement learning. *The International Journal of Robotics Research*, 42(10):827–858, 2023.

Wang, Y., Liu, L., and Dauwels, J. Slot-vae: Object-centric scene generation with slot attention. *ArXiv*, abs/2306.06997, 2023. URL https://api.semanticscholar.org/CorpusID:259138897.

Wang, Z., Gui, L., Negrea, J., and Veitch, V. Concept algebra for (score-based) text-controlled generative models. *Advances in Neural Information Processing Systems*, 36, 2024.

Wiedemer, T., Mayilvahanan, P., Bethge, M., and Brendel, W. Compositional generalization from first principles. *Advances in Neural Information Processing Systems*, 36, 2024.

Wu, T., Maruyama, T., Wei, L., Zhang, T., Du, Y., Iaccarino, G., and Leskovec, J. Compositional generative inverse design. *arXiv preprint arXiv:2401.13171*, 2024.

Yang, M., Du, Y., Dai, B., Schuurmans, D., Tenenbaum, J. B., and Abbeel, P. Probabilistic adaptation of text-to-video models. *arXiv preprint arXiv:2306.01872*, 2023a.

Yang, Z., Mao, J., Du, Y., Wu, J., Tenenbaum, J. B., Lozano-Pérez, T., and Kaelbling, L. P. Compositional diffusion-based continuous constraint solvers. *arXiv preprint arXiv:2309.00966*, 2023b.

Zhang, J., Liu, J., He, J., et al. Composing parameter-efficient modules with arithmetic operation. *Advances in Neural Information Processing Systems*, 36:12589–12610, 2023a.

Zhang, L., Rao, A., and Agrawala, M. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023b.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017.

A. Additional Related Works

Structured compositional generative models. Structured generative models leverage architectural inductive biases in an encoder-decoder framework, such as recurrent attention mechanisms (Gregor et al., 2015) or slot-attention (Wang et al., 2023). These models decompose scenes into background and parts-based representations in an unsupervised manner guided by modeling priors. While these approaches can flexibly generate scenes with single or multiple objects, they are not explicitly controllable, and require specific model pre-training on datasets containing compositions of interest.

Controllable generation. Composition at inference-time is one potential mechanism for exerting control over the generation process. Another way to modify compositions of style and/or content attributes is through spatial conditioning a pre-trained diffusion model on a structural attribute (e.g., pose or depth) as in Zhang et al. (2023b), or on multiple attributes of style and/or content as in Stracke et al. (2024). Another option is control through resampling, as in Liu et al. (2024). These methods are complementary to single or multiple model conditioning mechanisms based on score composition that we study in the current work.

Single model conditioning. We distinguish the kind of composition we study in this paper from approaches that rely on a single model but use OOD conditioners to achieve novel combinations of concepts never seen together during training; for example, passing OOD text prompts to text-to-image models (Nichol et al., 2021; Podell et al., 2023), or works like Okawa et al. (2024); Park et al. (2024) where a single model conditions simultaneously on multiple attributes like shape and color, with some combinations held out during training. In contrast, the compositions we study recombine the outputs of multiple separate models at inference time. Though less powerful, this can still be surprisingly effective, and is more amenable to theoretical study since it disentangles the potential role of conditional embeddings.

Multiple model composition. Among compositions involving multiple separate models, many different variants have been explored with different goals and applications. Some definitions of composition are inspired by logical operators like AND and OR, usually taken to mean that the composed distribution should have high probability under all of the conditional distributions to be composed, or at least one of them, respectively. Given two conditional probabilities $p_0(x), p_1(x)$, AND is typically implemented as the product $p_0(x)p_1(x)$ and OR as sum $p_0(x) + p_1(x)$ (though these only loosely correspond to the logical operators and other implementations are also possible). Some composition methods are based on diffusion models and use the learned scores (mainly for product compositions), others use energy-based models (which allows for OR-inspired sum compositions, as well as more sophisticated samplers, in particular sampling at $t = 0$ (Du et al., 2020; 2023; Liu et al., 2021), and still others work directly with the densities (Skreta et al., 2024) (enabling an even greater variety of compositions, including a different style of AND, taken to mean $p_0(x) = p_1(x)$). McAllister et al. (2025) explore another type of OR composition. (Wiedemer et al., 2024) take a different approach of taking the final rendered images generated by separate diffusion models and "adding them up" in pixel-space, as part of a study on generalization of data-generating processes. Task-arithmetic (Zhang et al., 2023a; Ilharco et al., 2022), often using LoRAs (Hu et al., 2021), is a kind of composition in weight-space that has had significant practical impact.

Product compositions. In this work, we focus specifically on product compositions (broadly defined to allow for a "background" distribution, i.e. compositions of the form $\hat{p}(x) = p_b(x) \prod_i \frac{p_i(x)}{p_b(x)}$) implemented with diffusion models, which allows the composition to be implemented via a linear combinations of scores as in Du et al. (2023); Liu et al. (2022). Our goal is not to propose a wholly new method of composition but rather to improve theoretical understanding of existing methods.

Learning and Generalization. Recently, Kamb & Ganguli (2024) demonstrated how a type of compositional generalization arises from inductive bias in the learning procedure (equivariance and locality). Their findings are relevant to our broader motivation, but complementary to the focus of this work. Specifically, we focus only on mathematical aspects of defining and sampling from compositional distributions, and we do not consider any learning-theoretic aspects such as inductive bias or sample complexity. This allows us to study the behavior of compositional sampling methods even assuming perfect knowledge of the underlying distributions.

B. CLEVR Experimental Details

All of our CLEVR experiments use raw conditional diffusion scores, without applying any guidance/CFG (Ho & Salimans, 2022). Details below.

B.1. Dataset, models, and training details

B.1.1. CLEVR DATASET

We used the CLEVR (Johnson et al., 2017) dataset generation procedure[6] to generate datasets customized to the needs of the present work. All default objects, shapes, sizes, colors were kept unchanged. Images were generated in their original resolution of 320×240 and down-sampled to a lower resolution of 128×128 to facilitate experimentation and to be more GPU resources friendly. The various datasets we generated from this procedure include:

- A background dataset (0 objects) with 50,000 samples

- Single object dataset with 1,550,000 samples

- A dataset having 1 to 5 objects, with 500,000 samples for each object count, for a total of 2,500,000 samples.

B.1.2. MODEL ARCHITECTURE

We used our own PyTorch re-implementation of the EDM2 (Karras et al., 2024) U-net architecture. Our re-implementation is functionally equivalent, and only differs in optimizations introduced to save memory and GPU cycles. We used the smallest model architecture, e.g. `edm2-img64-xs` from `https://github.com/NVlabs/edm2`. This model has a base channel width of 128, resulting in a total of 124M trainable weights. Two versions of this model were used:

- An unmodified version for background and class-conditioned experiments.

- A modified version for (x, y) conditioning in which we simply replaced Fourier embeddings for the class with concatenated Fourier embeddings for x and y.

B.1.3. TRAINING AND INFERENCE

In all experiments, the model is trained with a batch size of 2048 over 128×2^{20} samples by looping over the dataset as often as needed to reach that number. In practice, training takes around 16 hours to complete on 32 A100 GPUs. We used almost the same training procedure as in EDM2 (Karras et al., 2024), which is basically a standard training loop with gradient accumulation. The only difference is that we do weight renormalization after the weights are updated rather than before as the authors originally did.

For simplicity, we did not use posthoc-EMA to obtain the final weights used in inference. Instead we took the average of weights over the last 4096 training updates. The denoising procedure for inference is exactly the same as in EDM2 (Karras et al., 2024), e.g. 65 model calls using a 32-step Heun sampler.

B.2. Factorized Conditionals in CLEVR.

B.2.1. SINGLE OBJECT DISTRIBUTIONS WITH EMPTY BACKGROUND

Let us explicitly describe how our definition of Factorized Conditionals captures the CLEVR setting of Figures 2 and 3a. Recall, the background distribution p_b over n pixels is images of an empty scene with no objects. For each $i \in \{1, \dots, L\}$ (where $L = 4$ in Figure 2 and $L = 9$ in Figure 3(a)) define the set $M_i \in [n]$ as the set of pixel indices surrounding location i. Each M_i should be thought of as a "mask" that that masks out objects at location i. Then, let $M_b := (\cup_i M_i)^c$ be the remaining pixels in the image, excluding all the masks. Now we claim the distributions $(p_b, p_1, \dots, p_L)$ are approximately Factorized Conditionals, with corresponding coordinate partition $(M_b, M_1, \dots, M_L)$. We can confirm each criterion in Definition 5.2 individually:

---

[6] `https://github.com/facebookresearch/clevr-dataset-gen`

1. In each distribution p_i, the pixels inside the mask M_i are approximately independent from the pixels outside the mask, since the outside pixels always describe an empty scene.

2. In the background p_b, the set of masks $\{M_i\}$ specify approximately mutually-independent sets of pixels, since all pixels are roughly constant.

3. The distribution of p_i and p_b approximately agree along all pixels outside mask M_i, since they both describe an empty scene outside this mask.

Thus, the set of distributions approximately form Factorized Conditionals. However the conditions of Definition 5.2 do not *exactly* hold, since objects can cast shadows on each other and may even occlude each other. Empirically, this can significantly affect the results when composing many objects, as explored in Figure 3(a).

B.2.2. CLUTTERED DISTRIBUTED WITH UNCONDITIONAL BACKGROUND
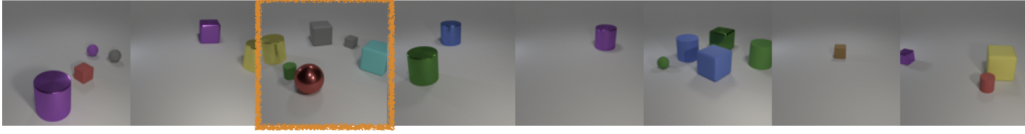

Figure 8: Samples from unconditional model trained on images containing 1-5 objects. The sampled images sometimes contain 6 objects (circled in orange).

Next, we discuss the setting of Figure 3c, which is a Bayes composition based on an unconditional distribution where each scene contains 1-5 objects (with the number of objects sampled uniformly). The locations and all other attributes of the objects are sampled independently. The conditions label the location of one randomly-chosen object. Just as in the previous case, for each $i \in \{1, 2, \dots, L\}$ ($L = 9$ in Figure 3c), we define the set $M_i \in [n]$ as the set of pixel indices surrounding location i, and let $M_b := (\cup_i M_i)^c$ be the remaining pixels in the image, excluding all the masks. Again, we claim that the distributions $(p_b, p_1, \dots, p_L)$ are approximately Factorized Conditionals, with corresponding coordinate partition $(M_b, M_1, \dots, M_L)$. We examine the criteria in Definition 5.2:

1. In each distribution p_i, the pixels inside the mask M_i are approximately independent from the pixels outside the mask, since the outside pixels approximately describe a distribution containing 0-4 objects, and the locations and other attributes of all objects are independent.

2. In the unconditional background distribution p_b, we argue that in practice, the set of masks $\{M_i\}$ are approximately mutually-independent. By assumption, the locations and other attributes of all shapes are all independent, and the masks M_i are chosen in these experiment to minimize interaction/overlap. The main difficulty is the restriction to 1-5 total objects, which we discuss below.

3. The distribution of p_i and p_b approximately agree along all pixels outside mask M_i, since $p_i|_{M_i^c}$ contains 0-4 objects, while $p_b|_{M_i^c}$ contains 0-5 objects (since one object could be 'hidden' in M_i^c).

There are, however, two important caveats to the points above. First, overlap or other interaction (shadows, etc.) between objects can clearly violate all three criteria. In our experiment, this is mitigated by the fact that the masks M_i are chosen to minimize interaction/overlap (though interactions start to occur as we compose more objects, leading to some image degradation). Second, since the number of objects is sampled uniformly from 1-5, the presence of one object affects the probability that another will be present. Thus, the masks $\{M_i\}$ are not perfectly independent under the background distribution, nor do p_i and p_b perfectly agree on M_i^c. Ideally, each p_i would place an object in mask M_i and independently follow p_b on M_i^c, and p_b would be such that the probability that an object appears in mask M_i is independently Bernoulli (c.f. Appendix E.2). In particular, this would imply that the distribution of the total number of objects is Binomial (which allows the total object-count to range from zero to the total-number-of-locations, as well as placing specific probabilities on each object-count), which clearly differs from the uniform distribution over 1-5 objects. However, a few factors mitigate this discrepancy:

- A Binomial with sufficiently small probability-of-success places very little probability on large k. For example, under Binomial$(9, 0.3)$, $\mathbb{P}(k = 0 : 5) = 0.04, 0.156, 0.27, 0.27, 0.17, 0.07$ and $\mathbb{P}(k > 5) = 0.026$.

- Empirically, the *learned* unconditional distribution does not actually enforce $k < 5$; we sometimes see samples with $k = 6$ for example, as seen in Figure 8.

Intuitively, the train distribution is "close to Bernoulli" and the *learned* distribution seems to be even closer.

With these considerations in mind, we see that the set of distributions approximately – though imperfectly – form Factorized Conditionals. One advantage of this setting compared to the single-object setting is that the models can learn how multiple objects should interact and even overlap correctly, potentially making it easier compose nearby locations. We explore the length-generalization of this composition empirically in Figure 3c (note, however, that only compositions of more than 5 objects are actually OOD w.r.t. the distributions p_i in this case).

B.3. Additional CLEVR samples

In this section we provide additional non-cherrypicked samples of the experiments shown in the main text.
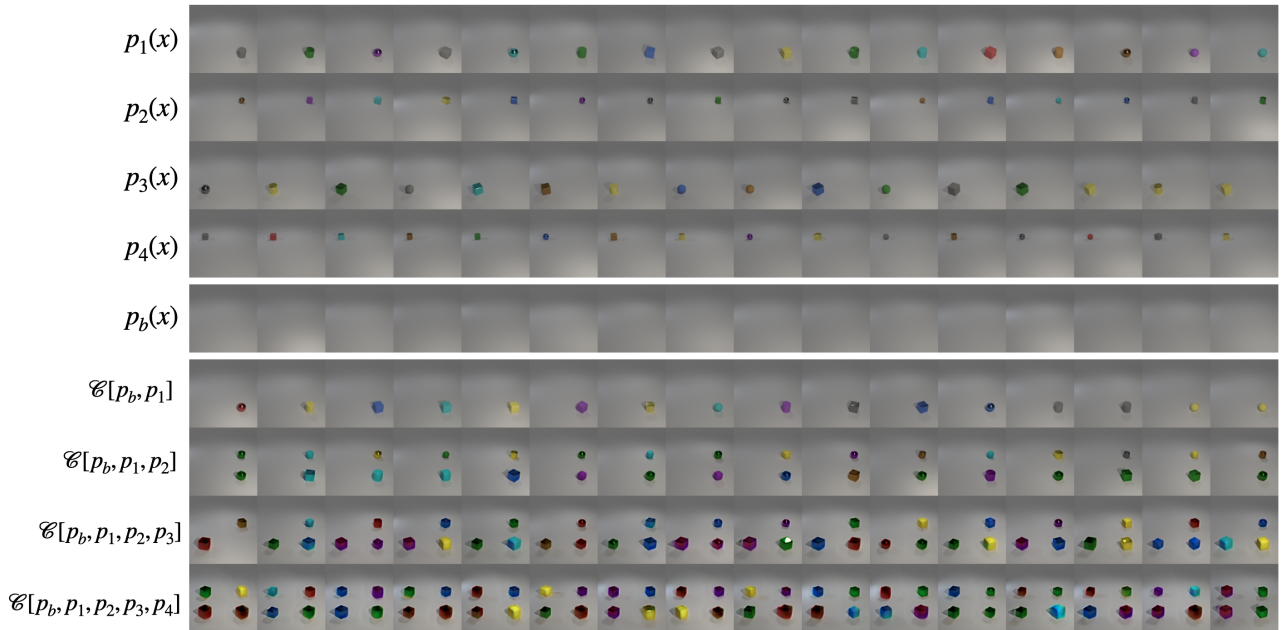

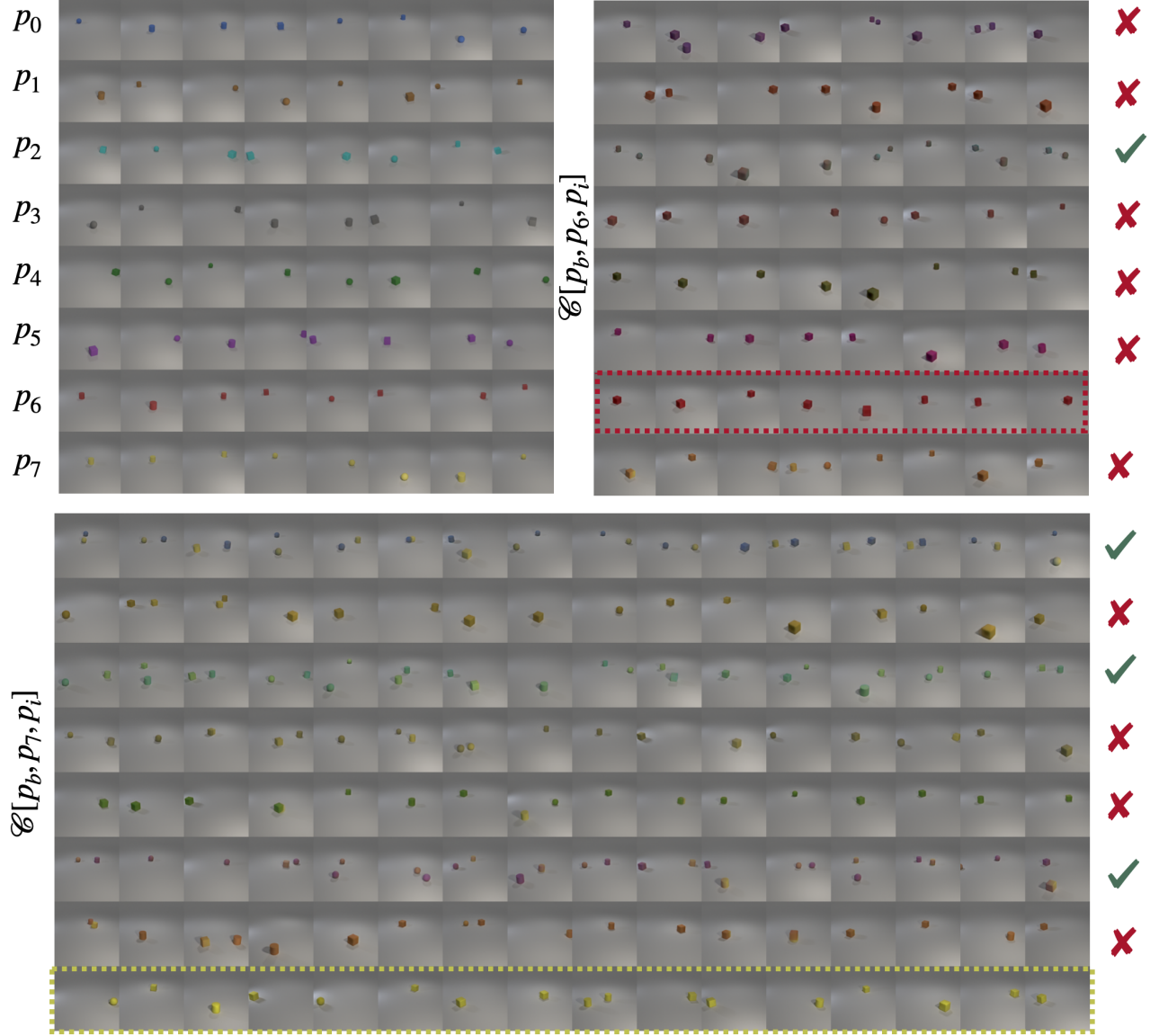Figure 9: Additional non-cherrypicked samples for CLEVR experiment of Figure 2.

Figure 10: Additional non-cherrypicked samples for CLEVR experiment of Figure 5. Top left grid shows conditional samples for each color. Top right grid shows compositions of red-colored objects (p_6) with objects of other colors (8 samples of each), which only succeeds for cyan-colored objects. Bottom grid shows compositions of yellow-colored objects (p_7) with objects of other colors (16 samples of each): these are additional samples of the exact experiment shown in Figure 5.

C. SDXL experimental details

C.1. Figure 1

The two models composed are

1. An SDXL model (Podell et al., 2023) fine-tuned on 30 personal photos of the author's dog (Papaya).

2. SDXL-base-1.0 (Podell et al., 2023) conditioned on prompt "an oil painting in the style of van gogh."

The background score distribution is the unconditional background (i.e. SDXL conditioned on the empty prompt). We use the DDPM sampler (Ho et al., 2020) with 30 steps, using the composed score, and CFG guidance weight of 2 Ho et al. (2020).

Note that using guidance weight 1 (i.e. no guidance) also performs reasonably in this case, but is lower quality.

C.2. Figure 7

Left: The two score models composed are

1. SDXL-base-1.0 (Podell et al., 2023) conditioned on prompt "photo of a dog"

2. SDXL-base-1.0 (Podell et al., 2023) conditioned on prompt "photo of a horse"

The background score distribution is the unconditional background (i.e. SDXL conditioned on the empty prompt).

For improved sample quality, we use a Predictor-Corrector method (Song et al., 2020) with the DDPM predictor and the Langevin dynamics corrector, both operating on the composed score. We use 100 predictor denoising steps, and 3 Langevin iterations per step. We do not use any guidance/CFG.

Right: Identical setting as above, using prompts:

1. "photo of a dog"

2. "photo, with red hat"

Note that the DDPM sampler also performed reasonably in this setting, but Predictor-Corrector methods improved quality.

D. Reverse Diffusion and other Samplers

D.1. Diffusion Samplers

DDPM (Ho et al., 2020) and DDIM (Song et al., 2021) are standard reverse diffusion samplers (Sohl-Dickstein et al., 2015; Song & Ermon, 2019) that correspond to discretizations of a reverse-SDE and reverse-ODE, respectively (so we will sometimes refer to the reverse-SDE as DDPM and the reverse-ODE as DDIM for short). The forward process, reverse-SDE, and equivalent reverse-ODE (Song et al., 2020) for the *variance-preserving* (VP) (Ho et al., 2020) conditional diffusion are

$$\text{Forward SDE}: dx = -\frac{1}{2}\beta_t x dt + \sqrt{\beta_t} dw. \tag{12}$$

$$\text{DDPM SDE}: \quad dx = -\frac{1}{2}\beta_t x\, dt - \beta_t \nabla_x \log p_t(x|c) dt + \sqrt{\beta_t} d\bar{w} \tag{13}$$

$$\text{DDIM ODE}: \quad dx = -\frac{1}{2}\beta_t x\, dt - \frac{1}{2}\beta_t \nabla_x \log p_t(x|c) dt. \tag{14}$$

D.2. Langevin Dynamics

Langevin dynamics (LD) (Rossky et al., 1978; Parisi, 1981) an MCMC method for sampling from a desired distribution. It is given by the following SDE (Robert et al., 1999)

$$dx = \frac{\varepsilon}{2}\nabla \log \rho(x) dt + \sqrt{\varepsilon} dw, \tag{15}$$

which converges (under some assumptions) to $\rho(x)$ (Roberts & Tweedie, 1996). That is, letting $\rho_s(x)$ denote the solution of LD at time s, we have $\lim_{s\to\infty} \rho_s(x) = \rho(x)$.

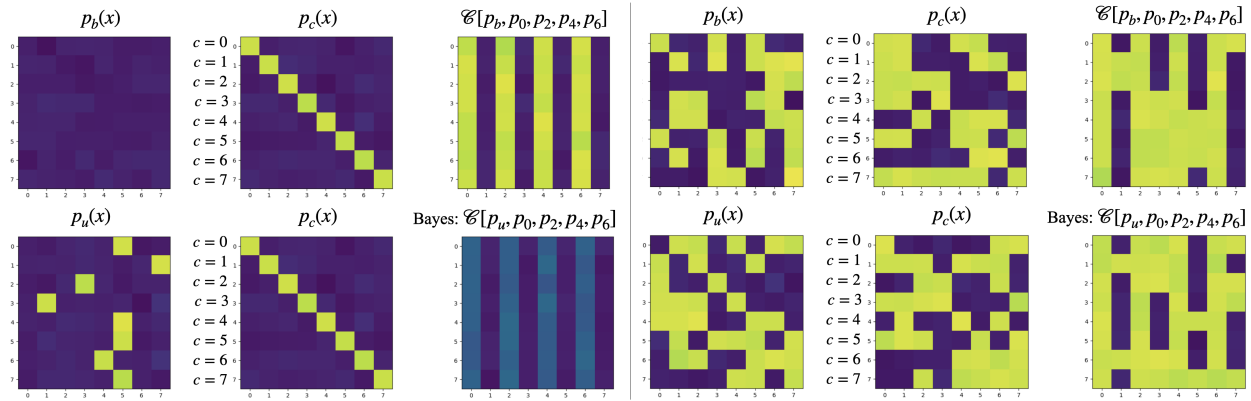# E. Connections with the Bayes composition

Figure 11: Bayes composition vs. projective composition. All experiments use exact scores, which is possible since the diffusion-noised distributions are Gaussian mixtures. (Left) Distributions follow (16): each conditional p_i activates index i only, unconditional p_u averages over the p_i, and background p_b is all-zeros. We attempt to compose the conditions p_0, p_2, p_4, p_6 and hope to obtain the result $[1, 0, 1, 0, 1, 0]$. This requires length-generalization, since each of the conditionals p_i contains only a single 1. The composition using the empty background p_b (top) achieves this goal, while the Bayes composition using the unconditional p_u (bottom) does not. Note that $[p_b, p_1, p_2, \ldots]$ satisfy Definition 5.2 while $[p_u, p_1, p_2, \ldots]$ does not. (Right) Distributions follow (17), where each conditional p_i activates index i on an independently 'cluttered' background. In this case the unconditional is similar to the cluttered background. Again we attempt to compose p_0, p_2, p_4, p_6, and in this case we find that the composition using p_u works similarly well to p_b.

E.1. The Bayes composition and length-generalization

We give a counterexample for which the Bayes composition fails to length-generalize, while composition using an "empty background" succeeds. The example corresponds to the experiment shown in Figure 11 (left). Suppose we have conditional

distributions p_i that set a single index i to one and all other indices to zero, a zero-background distribution p_b, and an unconditional distribution formed from the conditionals by assuming $p(c = i)$ is uniform. That is:

$$p_i^t(x_t) = \mathcal{N}(x_t; e_i, \sigma_t^2) \propto \exp\left(-\frac{\|x_t - e_i\|^2}{2\sigma_t^2}\right)$$

$$p_b^t(x_t) = \mathcal{N}(x_t; 0, \sigma_t^2) \propto \exp\left(-\frac{\|x_t\|^2}{2\sigma_t^2}\right)$$

$$p_u^t(x_t) = \frac{1}{n}\sum_{i=1}^n p_i(x_t) \tag{16}$$

Suppose we want to compose all n distributions p_i, that is, we want to activate all indices. It is enough to consider x_t of the special form $x_t = (\alpha, \dots, \alpha)$ since there is no reason to favor any condition over any another. Making this restriction,

$$x_t = (\alpha, \dots, \alpha) \implies p_i^t(x_t) \propto \exp\left(-\frac{(n-1)\alpha^2 + (1-\alpha)^2}{2\sigma_t^2}\right) = \exp\left(-\frac{n\alpha^2 - 2\alpha + 1}{2\sigma_t^2}\right), \quad \forall i$$

$$p_u^t(x_t) = \exp\left(-\frac{n\alpha^2 - 2\alpha + 1}{2\sigma_t^2}\right)$$

$$p_b^t(x_t) \propto \exp\left(-\frac{n\alpha^2}{2\sigma_t^2}\right)$$

Let us find the value of α that maximizes the probability under the Bayes composition of all condition:

$$x_t = (\alpha, \dots, \alpha) \implies \frac{p_i^t(x_t)}{p_u^t(x_t)} = 1$$

$$\implies p_u^t(x_t) \prod_{i=1}^n \frac{p_i^t(x_t)}{p_u^t(x_t)} \propto p_u^t(x_t) \propto \exp\left(-\frac{n\alpha^2 - 2\alpha + 1}{2\sigma_t^2}\right) = \exp\left(-\frac{n(\alpha - \frac{1}{n})^2 + \text{const}}{2\sigma_t^2}\right)$$

$$\implies \alpha^\star = \frac{1}{n},$$

so the optimum is $\alpha^\star = \frac{1}{n}$. That is, under the Bayes composition the most likely configuration places value $\frac{1}{n}$ at each index we wished to activate, rather than the desired value 1.

On the other hand, if we instead use p_b in the linear score combination and optimize, we find that:

$$x_t = (\alpha, \dots, \alpha) \implies \implies \frac{p_i^t(x_t)}{p_b^t(x_t)} \propto \exp\left(-\frac{1 - 2\alpha}{2\sigma_t^2}\right)$$

$$\implies p_b^t(x_t) \prod_{i=1}^n \frac{p_i^t(x_t)}{p_b^t(x_t)} \propto \exp\left(-\frac{n\alpha^2}{2\sigma_t^2}\right)\exp\left(-\frac{n(1 - 2\alpha)}{2\sigma_t^2}\right) \propto \exp\left(-\frac{n(\alpha^2 - 2\alpha + 1)}{2\sigma_t^2}\right)$$

$$\propto \exp\left(-\frac{n(\alpha - 1)^2}{2\sigma_t^2}\right)$$

$$\implies \alpha^\star = 1$$

so the optimum is $\alpha^\star = 1$. That is, the most likely configuration places the desired value 1 at each index we wished to activate, achieving projective composition, and in particular, length-generalizing correctly.

E.2. Cluttered Distributions

In certain "cluttered" settings, the Bayes composition may be approximately projective. We explore this in the following simplified setting, corresponding to the experiment in Figure 11 (right). Suppose that x is binary-valued, $M_i = \{i\}, \forall i$, the x_i are independently Bernoulli with parameter q under the background, and the projected conditional distribution $p_i(x|_i)$ just guarantees that $x_i = 1$:

$$p_b(x|_{i^c}) \sim \text{Bern}_q(x|_{i^c}), \text{ i.i.d. } \forall i, \qquad p_i(x|_i) = \mathbb{1}_{x|_i = 1}, \tag{17}$$

The distributions $(p_b, p_1, p_2, \ldots)$ then clearly satisfy Definition 5.2 and hence guarantee projective composition. In this case, the unconditional distribution used in the Bayes composition is similar to the background distribution if number of conditions is large. Intuitively, each conditional looks very similar to the Bernoulli background except for a single index that is guaranteed to be equal to 1, and the unconditional distribution is just a weighted sum of conditionals. Therefore, we expect the Bayes composition to be approximately projective.

More precisely, we will show that the unconditional distribution converges to the background in the limit as $n \to \infty$, where n is both the data dimension and number of conditions, in the following sense:

$$\mathop{\mathbb{E}}_{x \sim p_b} \left[\left(\frac{p_u(x) - p_b(x)}{p_b(x)} \right)^2 \right] \to 0 \quad \text{as } n \to \infty.$$

We define the conditional and background distributions by:

$$x \in \mathbb{R}^n, \quad M_i = \{i\}$$
$$p_b(x|_i) \sim \text{Bern}_q(x|_i), \text{ i.i.d. for } i = 1, \ldots, n$$
$$p_i(x|_i) = \mathbb{1}_{x|_i=1}, \text{ for all } i = 1, \ldots, n$$
$$\implies p_b(x) = q^{nnz(x)}(1-q)^{n-nnz(x)}$$
$$p_i(x) = \mathbb{1}_{x|_i=1} p_b(x|_{i^c}) = \mathbb{1}_{x|_i=1} q^{nnz(x|_{i^c})}(1-q)^{n-1-nnz(x|_{i^c})}$$

We construct the unconditional distribution with assuming uniform probability over all labels: $p_u(x) := \frac{1}{n} \sum_i p_i(x)$. The number-of-nonzeros (nnz) in all of these distributions follow Binomial distributions:

$$x \sim p_b \implies p_b(nnz(x) = k) \sim \text{Binom}(k; n, q)$$
$$x \sim p_i \implies p_i(nnz(x) = k) = p_b(nnz(x|_{i^c}) = k - 1)$$
$$\sim \text{Binom}(k - 1; n - 1, q) \quad \text{if } k > 0 \text{ else } 0$$
$$x \sim p_u \implies p_u(nnz(x) = k) = \frac{1}{n} \sum p_i(nnz(x) = k)$$
$$\sim \text{Binom}(k - 1; n - 1, q) \quad \text{if } k > 0 \text{ else } 0$$

The basic intuition is that for large k and n, $p_b \sim \text{Binom}(k; n, q)$ and $p_u \sim \text{Binom}(k - 1; n - 1, q)$ are similar. More precisely, we can calculate:

$$\mathop{\mathbb{E}}_{x \sim p_b} \left[\left(\frac{p_u(x) - p_b(x)}{p_b(x)} \right)^2 \right] = \mathop{\mathbb{E}}_{x \sim p_b} \left[\left(\frac{nnz(x)}{qn} - 1 \right)^2 \right], \quad \text{since } \frac{B(k - 1; n - 1, q)}{B(k; n, q)} = \frac{k}{qn}$$

$$= \mathop{\mathbb{E}}_{k \sim \text{Binom}(n,q)} \left[\left(\frac{k}{qn} - 1 \right)^2 \right] = \frac{1}{(nq)^2} \mathop{\mathbb{E}}_{k \sim \text{Binom}(n,q)} \left[(k - nq)^2 \right]$$

$$= \frac{1}{(nq)^2} \text{Var}(k), \quad k \sim \text{Binom}(n, q)$$

$$= \frac{1}{(nq)^2} nq(1-q) = \frac{1-q}{nq} \to 0 \quad \text{as } n \to \infty.$$

F. Factorized conditionals vs. orthogonal score differences

To see that Definition 5.2 implies orthogonality between the score differences, we note that

$$v_i^t(x) := \nabla_x \log p_i^t(x_t) - \nabla_x \log p_b^t(x_t)$$
$$= \nabla_x \log \frac{p_i^t(x)}{p_b^t(x)} = \nabla_x \log \frac{p_i^t(x|_{M_i}) p_b^t(x|_{M_i^c} x)}{p_b^t(x|_{M_i}) p_b^t(x|_{M_i^c})}$$
$$= \nabla_x \log \frac{p_i^t(x|_{M_i})}{p_b^t(x|_{M_i})}$$
$$\implies v_i^t(x)[k] = 0, \quad \forall k \notin M_i$$
$$\implies v_i^t(x)^T v_j^t(x) = 0, \quad \forall i \neq j, \quad \text{since } M_i \cap M_j = \emptyset,$$

where in the second-to-last line we used the fact that the gradient of a function depending only on a subset of variables has zero entries in the coordinates outside that subset.

In fact, the same argument implies that $\{v_i^t(x) : x \in \mathbb{R}^n\} \subset M_i$; in other words, $\{v_i^t(x) : x \in \mathbb{R}^n\}$ and $\{v_j^t(x) : x \in \mathbb{R}^n\}$ occupy mutually-orthogonal subspaces. But even this latter condition does not imply the stronger condition of Definition 5.2. To find an equivalent definition in terms of scores we must also capture the independence of the subsets under p_b. Specifically:

$$
\begin{cases}
p_i^t(x) = p_i^t(x|_{M_i}x)p_b^t(x|_{M_i^c}x) \\
p_b^t(x) = p_b^t(x|_{\bar{M}}x)\prod_i p_b^t(x|_{M_i})
\end{cases}
$$

$$
\iff
\begin{cases}
\nabla_x \log p_i^t(x) = \nabla_x \log p_i^t(x|_{M_i}x) + \nabla_x \log p_b^t(x|_{M_i^c}x) \\
\nabla_x \log p_b^t(x) = \nabla_x \log p_b^t(x|_{\bar{M}}x) + \sum_i \nabla_x \log p_b^t(x|_{M_i})
\end{cases}
$$

$$
\iff
\begin{cases}
\nabla_x \log p_i^t(x) - \nabla_x \log p_b^t(x) = \nabla_x \log \dfrac{p_i^t(x|_{M_i}x)}{p_b^t(x|_{M_i}x)} \\
\nabla_x \log p_b^t(x) = \nabla_x \log p_b^t(x|_{\bar{M}}x) + \sum_i \nabla_x \log p_b^t(x|_{M_i})
\end{cases}
$$

So an equivalent definition in terms of scores could be:

Definition F.1. *The distributions* $(p_b, p_1, p_2, \ldots)$ *form* factored conditionals *if the score-deltas* $v_i^t := \nabla_x \log p_i^t(x) - \nabla_x \log p_b^t(x)$ *satisfy* $\{v_i^t(x) : x \in \mathbb{R}^n\} \subset M_i$, *where the* M_i *are mutually-orthogonal subsets, and furthermore the score of the background distribution decomposes over these subsets as follows:* $\nabla_x \log p_b^t(x) = \nabla_x \log p_b^t(x|_{\bar{M}}x) + \sum_i \nabla_x \log p_b^t(x|_{M_i})$.

(Note: this is actually equivalent to a slightly more general version of Definition 5.2 that allows for orthogonal transformations, which is the most general assumption under which diffusion sampling generates a projective composition, per Lemmas 6.1 and 7.1.)

G. Proof of Theorem 5.3

Proof. (Theorem 5.3) For any set of distributions $\vec{q} = (q_b, q_1, q_2, \ldots)$ satisfying Definition 5.2, we have

$$\mathcal{C}[\vec{q}](x) := q_b(x) \prod_i \frac{q_i(x)}{q_b(x)} = q_b(x) \prod_i \frac{q_b(x_t|_{M_i^c})q_i(x|_{M_i})}{q_b(x|_{M_i^c})q_b(x|_{M_i})}$$

$$= q_b(x) \prod_i \frac{q_i(x|_{M_i})}{q_b(x|_{M_i})} = q_b(x|_{M_b}) \prod_i q_i(x_t|_{M_i}) \qquad (18)$$

(where we used (7) in the second equality). Since $(p_b, p_1, p_2, \ldots)$ satisfy Definition 5.2 by assumption, applying (18) gives

$$\mathcal{C}[\vec{p}](x) = p_b(x|_{M_b}) \prod_i p_i(x|_{M_i}) := \hat{p}(x),$$

so the composition at $t = 0$ is projective, as desired. Now to show that reverse-diffusion sampling with the compositional scores generates $\mathcal{C}[\vec{p}]$, we need to show that

$$\mathcal{C}[\vec{p^t}] = N_t[\mathcal{C}[\vec{p}]],$$

where $p^t := N_t[p]$ denotes the t-noisy version of distribution p under the forward diffusion process. First, notice that if $\vec{p}$ satisfies Definition 5.2, then $\vec{p^t}$ does as well. This is because the diffusion process adds Gaussian noise independently to each coordinate, and thus preserves independence between sets of coordinates. Therefore by (18), we have $\mathcal{C}[\vec{p^t}](x) = p_b^t(x|_{\bar{M}}) \prod_i p_i^t(x_t|_{M_i})$. Now we apply the same argument (that diffusion preserves independent sets of coordinates) once again, to see that $\mathcal{C}[\vec{p^t}] = N_t[\mathcal{C}[\vec{p}]]$, as desired. $\qquad \square$

H. Parameterization-Independent Compositions and Proof of Lemma 6.1

The proof of Lemma 6.1 relies on certain general fact about parametrization-independence of certain operators, which we develop here.

Suppose we have an operator that takes as input two probability distributions (p, q) over the same space $\mathcal{X}$, and outputs a distribution over $\mathcal{X}$. That is, $F : \Delta(\mathcal{X}) \times \Delta(\mathcal{X}) \to \Delta(\mathcal{X})$. We can think of such operators as performing some kind of "composition" of p, q.

Certain operators are *independent of parameterization*, meaning for any reparameterization of the base space $A : \mathcal{X} \to \mathcal{Y}$, we have

$$F(p, q) = A^{-1}\sharp(F(A\sharp p, A\sharp q))$$

or equivalently:

$$F(A\sharp p, A\sharp q) = A\sharp F(p, q),$$

where $\sharp$ is the pushforward:

$$(\mathcal{A}\sharp p)(z) := \frac{1}{|\nabla \mathcal{A}|} p(\mathcal{A}^{-1}(z)).$$

This means that reparameterization commutes with the operator: it does not matter if we first reparameterize, then compose, or first compose, then reparamterize. A few examples:

1. The pointwise-geometric median, $F(p, q)(x) := \sqrt{p(x)q(x)}$, is independent of reparameterization:

2. Squaring a distribution, $F(p, q)(x) := p(x)^2$, is NOT independent of reparameterization:

3. The "CFG composition" (Ho & Salimans, 2022), $F(p, q)(x) := p(x)^\gamma q(x)^{1-\gamma}$, is independent of reparameterization:

We can analogously define parametrization-independence for operators on more than 2 distributions. Notably, given a tuple of distributions $\vec{p} = (p_b, p_1, p_2, \ldots, p_k)$, our composition operator $\mathcal{C}$ of Definition 5.1, $\mathcal{C}[\vec{p}] \propto p_b(x) \prod_i \frac{p_i(x)}{p_b(x)}$ is independent of parameterization.

Lemma H.1 (Parametrization-independence of 1-homogeneous operators)**.** *If an operator F is 1-homogeneous, i.e. $F(tp, tq, \ldots) = tF(p, q, \ldots)$ and operates pointwise, then it is independence of parametrization.*

Proof.

$$
\begin{aligned}
F(\mathcal{A}\sharp p, \mathcal{A}\sharp q, \ldots)(z) &= F(\mathcal{A}\sharp p(z), \mathcal{A}\sharp q(z), \ldots), \quad \text{pointwise} \\
&= F\left(\frac{1}{|\nabla\mathcal{A}|} p(\mathcal{A}^{-1}(z)), \frac{1}{|\nabla\mathcal{A}|} q(\mathcal{A}^{-1}(z)), \ldots \right) \\
&= \frac{1}{|\nabla\mathcal{A}|} F\left(p(\mathcal{A}^{-1}(z)), q(\mathcal{A}^{-1}(z)), \ldots \right), \quad \text{1-homogeneous} \\
&= \mathcal{A}\sharp F(p, q, \ldots)(z)
\end{aligned}
$$

$\square$

Corollary H.2 (Parametrization-invariance of composition)**.** *The composition operator $\mathcal{C}$ given by Definition 5.1 is independent of parametrization.*

Proof. The composition operator given by Definition 5.1 is 1-homogeneous:

$$
\mathcal{C}(tp_b, tp_1, tp_2, \ldots)(x) = tp_b(x) \prod_i \frac{tp_i(x)}{tp_b(x)} = tp_b(x) \prod_i \frac{p_i(x)}{p_b(x)} = t\mathcal{C}(p_b, p_1, p_2, \ldots)(x)
$$

and so the result follows from Lemma H.1. Alternatively, a direct proof is:

$$
\mathcal{C}(p_b, p_1, p_2, \ldots)(x) := p_b(x) \prod_i \frac{p_i(x)}{p_b(x)}
$$

$$
\mathcal{C}(\mathcal{A}\sharp p_b, \mathcal{A}\sharp p_1, \mathcal{A}\sharp p_2, \ldots)(z) = (\mathcal{A}\sharp p_b)(z) \prod_i \frac{(\mathcal{A}\sharp p_i)(z)}{(\mathcal{A}\sharp p_b)(z)} = \frac{1}{|\nabla\mathcal{A}|} p_b(\mathcal{A}^{-1}(z)) \prod_i \frac{p_i(\mathcal{A}^{-1}(z))}{p_b(\mathcal{A}^{-1}(z))} = \mathcal{A}\sharp \mathcal{C}(p_b, p_1, p_2, \ldots)(z).
$$

$\square$

Theorem 6.1 follows from Corollary H.2:

Proof. (Theorem 6.1) Let $(q_b, q_1, q_2, \ldots, q_k) := (\mathcal{A}\sharp p_b, \mathcal{A}\sharp p_1, \ldots \mathcal{A}\sharp p_k)$, for which Definition 5.2 holds by assumption. Applying an intermediate result from the proof of Theorem 5.3 gives:

$$
\mathcal{C}[\vec{q}](z) := q_b(z) \prod_i \frac{q_i(z)}{q_b(z)} = q_b(z|_{\bar{M}}) \prod_i q_i(z|_{M_i}).
$$

By Corollary H.2, $\mathcal{C}$ is independent of parametrization, hence

$$
\mathcal{A}\sharp \hat{p} := \mathcal{A}\sharp(\mathcal{C}[\vec{p}]) = \mathcal{C}[\mathcal{A}\vec{\sharp}p] := \mathcal{C}(\vec{q}).
$$

$\square$

I. Proof of Lemma 7.1

Figure 12 shows a synthetic experiment illustrating the sampling guarantees of Lemma 7.1 in contrast to the lack-of-guarantees in the non-orthogonal case.

The proof of Lemma 7.1 relies on the fact that diffusion noising commutes with orthogonal transformation, i.e. $\mathcal{A}\sharp N_t[q] = N_t[\mathcal{A}\sharp q]$ if $\mathcal{A}$ is orthogonal, since standard Gaussians are invariant under orthogonal transformation.
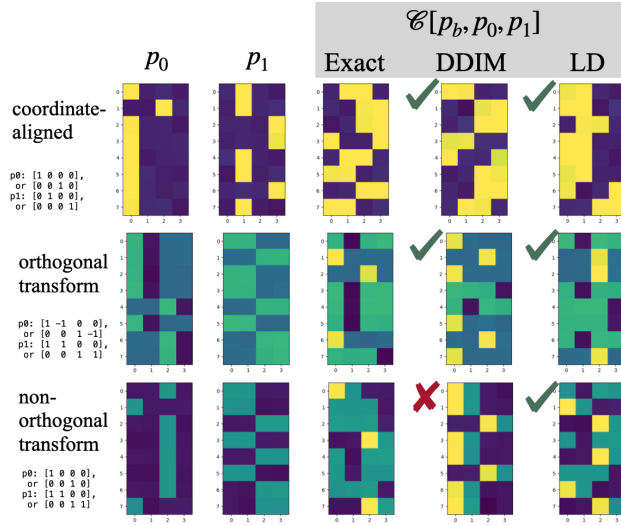
Figure 12: Synthetic composition experiment illustrating the sampling guarantees of Lemma 7.1 in contrast to the lack-of-guarantees in the non-orthogonal case. We compare a coordinate-aligned case (which satisfies Definition 5.2 in the native space) (top), an orthogonal-transform case (middle) (which satisfies the assumptions of Lemma 7.1), and a non-orthogonal-transform case (bottom) (which satisfies the assumptions of Theorem 6.1 but not of Lemma 7.1). In the first two cases the correct composition can be sampled using either diffusion (DDIM) or Langevin dynamics (LD) at $t = 0$, while in the final case DDIM sampling is unsuccessful although LD at $t = 0$ still works. The distributions are 4-dimensional and we show 8 samples (rows) for each. We show samples from the individual conditional distributions p_0, p_1 using DDIM, samples from the desired exact composition $\mathcal{C}[p_b, p_0, p_1]$ at $t = 0$ (obtained by sampling from $\mathcal{A}\sharp\mathcal{C}[\vec{p}]$ with DDIM and transforming by $\mathcal{A}^{-1}$), samples from the composition $\mathcal{C}[p_b, p_0, p_1]$ using DDIM with exact scores, and samples from the composition $\mathcal{C}[p_b, p_0, p_1]$ using Langevin dynamics (LD) with exact scores at time $t = 0$ in the diffusion schedule ($\sigma_{\min} = 0.02$). The noiseless distributions p_0 and p_1 are each 4-dimensional 2-cluster Gaussian mixtures with means as noted in the figure, equal weights, and standard deviation $\tau = 0.02$. For example, in the non-orthogonal-transform case, p_0 has means $[1, 0, 0, 0]$ and $[0, 0, 1, 0]$, and p_1 has means $[1, 1, 0, 0]$ and $[0, 0, 1, 1]$, (which can be transformed to satisfy Definition 5.2 via a non-orthogonal linear transform).

Proof. (Lemma 7.1) By assumption, $(\mathcal{A}\sharp p_b, \mathcal{A}\sharp p_1, \ldots \mathcal{A}\sharp p_k)$ satisfy Definition 5.2, where $\mathcal{A}(z) = Az$ with A an orthonormal matrix. By Lemma 6.1, $\hat{p} = \mathcal{C}[\vec{p}]$ satisfies (10). To show that reverse-diffusion sampling with scores $s_t = \nabla_x \log \mathcal{C}[\vec{p}^t]$ generates the composed distribution $\mathcal{C}[\vec{p}]$ we need to show that composition commutes with the forward diffusion process, i.e.

$$\mathcal{C}[\vec{p^t}] = N_t[\mathcal{C}[\vec{p}]].$$

Theorem 5.3 immediately gives us

$$\mathcal{C}[N_t[\mathcal{A}\sharp p]] = N_t[\mathcal{C}[\mathcal{A}\sharp p]].$$

Now we have to be careful with commuting operators. We know that composition is independent of parametrization, i.e. $\mathcal{A}\sharp\mathcal{C}[\vec{p}] = \mathcal{C}[\mathcal{A}\vec{\sharp}p]$. Diffusion noising N_t commutes with orthogonal transformation, i.e. $\mathcal{A}\sharp N_t[q] = N_t[\mathcal{A}\sharp q]$ if $\mathcal{A}$ is orthogonal, because a standard Gaussian multiplied by an orthonormal matrix Q remains a standard Gaussian: $\eta \sim \mathcal{N}(0, I) \implies Q\eta \sim \mathcal{N}(0, QQ^T) = \mathcal{N}(0, I)$ (this is false for non-orthogonal transforms, however). Therefore, in the orthogonal case, we can rewrite:

$$\mathcal{A}\sharp\mathcal{C}[N_t[p]] = \mathcal{A}\sharp N_t[\mathcal{C}[p]],$$

which implies the desired result since $\mathcal{A}$ is invertible. $\qquad\square$

24

J. Proof and further discussion of Lemma 7.2

J.1. Benefits of sampling at $t = 0$

Interestingly, (Du et al., 2023) have observed that sophisticated samplers like Hamiltonian Monte Carlo (HMC) requiring energy-based formulations often outperform standard diffusion sampling for compositional quality. Lemmas 6.1 and 7.2 help explain why this may be the case. In particular, HMC (or any variant of Langevin dynamics) can enable sampling p^0 at time $t = 0$, even when the path p^t used for annealing does not necessarily represent a valid forward diffusion process starting from p^0 (as Du et al. (2023) note, $\mathcal{C}[\vec{p}^t]]$ may not be). Lemma 6.1 should gives us hope that approximately-projective composition may often be possible at $t = 0$, since it allows *any* invertible transform $\mathcal{A}$ to transform into a factorized feature space (which need not be explicitly constructed). However, that does not mean that we can actually *sample* from this projection at time $t = 0$. As Lemma 7.2 shows, $\mathcal{C}[\vec{p}^t]]$ is not necessarily a valid diffusion path unless $\mathcal{A}$ is orthogonal, so standard diffusion sampling may not work. This is consistent with Du et al. (2023)'s observation that non-diffusion samplers that allow sampling at $t = 0$ may be necessary. Interestingly, Lemma 7.2 further cautions that sometimes $\mathcal{C}[\vec{p}^t]]$ may not even be an effective annealing path for any kind of sampler (which is consistent with our own experiments but not reported by other works, to our knowledge.)
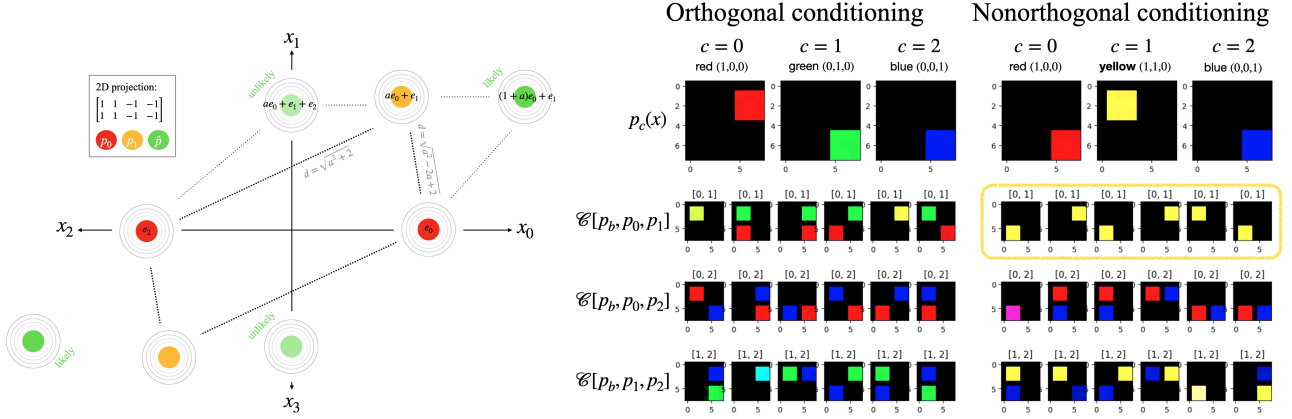
J.2. Proof of Lemma 7.2

Figure 13: (Left) A visualization of the intuition behind the proof of Lemma 7.2, under a 2D projection. (Right) An experiment where the colors red, green, and blue all compose projectively, while the colors red and yellow do not. We trained a Unet on images each containing a single square in one of 4 locations (selected randomly) and a certain color, conditioned on the color. We then generate composed distributions by running DDIM on the composed scores. The desired result of composing red and blue is an image containing a red and a blue square, both with randomly-chosen locations (so we occasionally get a purple square when the locations overlap). When we try to compose red and yellow, we only only ever obtain a single yellow square. Note that in pixel space, the colors are represented as red $(1, 0, 0)$, green $(0, 1, 0)$, blue $(0, 0, 1)$, yellow $(1, 1, 0)$, so that red, green and blue are all orthogonal and are expected to work by Lemma 5.3, while red and yellow are not orthogonal, and fail as allowed by Lemma 7.2. In fact this experiment is closely related to the counterexample used to prove Lemma 7.2.

We will prove Lemma 7.2 using a counterexample, which is inspired by an experiment, shown in Figure 14 (left), where non-orthogonal conditions fail to compose projectively.

The basic idea for the counterexample is that given a distribution $p(x)$ with two conditions, $c = 0, 1$, such at $t = 0$,

$$p_0(x) \approx \frac{1}{2}\delta_{e_0}(x) + \frac{1}{2}\delta_{e_2}(x), \qquad p_1(x) \approx \frac{1}{2}\delta_{ae_0+e_1}(x) + \frac{1}{2}\delta_{ae_2+e_3}(x),$$

for some $0 < a \le 1$, so the conditional distributions do not satisfy the independence assumption of Definition 5.2, However,
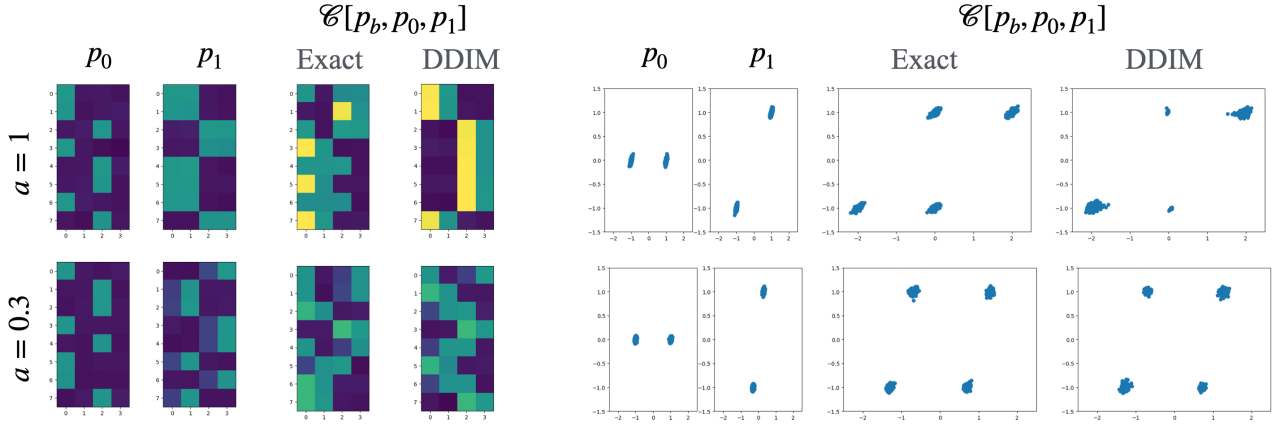
Figure 14: Composition experiments for the setting in the proof of Lemma 7.2. Left pane shows 8 samples (rows) of each distribution in the native 4d representation; right pane shows 1000 samples under the 2D projection used in Figure 13. We show samples from the individual conditional distributions p_0, p_1 using DDIM, samples from the desired exact composition $\mathcal{C}[p_b, p_0, p_1]$ at $t = 0$ (obtained by sampling from $\mathcal{A}\sharp\mathcal{C}[\vec{p}]$ with DDIM and transforming by $\mathcal{A}^{-1}$), and samples from the composition $\mathcal{C}[p_b, p_0, p_1]$ using DDIM with exact scores. We take $\tau = 0.02$, and set $\sigma_{\min} = 0.02$ in the diffusion schedule. In the top row we take $a = 1$ ("very non-orthogonal") as in the proof, and compare this to $a = 0.3$ ("mildly non-orthogonal") in the bottom row. With $a = 1$, as in the proof we see that DDIM barely samples two of the clusters. With $a = 0.3$, DDIM still slightly undersamples the "hard" clusters but the effect is much less pronounced.

there exists a (linear, but not orthogonal) A such that the distribution of $z = Ax$ is axis-aligned

$$(A\sharp p_0)(z) \approx \frac{1}{2}\delta_{e_0}(x) + \frac{1}{2}\delta_{e_2}(x), \qquad (A\sharp p_1)(z) \approx \frac{1}{2}\delta_{e_1}(x) + \frac{1}{2}\delta_{e_3}(x),$$

and thus does satisfy Definition 5.2 at $t = 0$, which guarantees correct composition of p at $t = 0$ under Lemma 6.1. The correct composition should sample uniformly from $\{(1+a)e_0+e_1, \quad e_0+ae_2+e_3, \quad ae_0+e_2+e_1, \quad (1+a)e_2+e_3\}$. What goes wrong is that as soon as we add Gaussian noise to the distribution $p(x)$ at time $t > 0$ of the diffusion forward process, the relationship $z = Ax$ breaks and so we are no longer guaranteed correct composition of $p^t(x)$. In fact, the distribution is still a GMM but places nearly all its weight on only two of the four clusters, namely: $\{(1 + a)e_0 + e_1, (1 + a)e_2 + e_3\}$. Intuitively, let us focus on the mode $ae_0 + e_1$ of p_1 and consider how it interacts with the two modes e_0, e_2 of p_0, at some time $t > 0$ when we have isotropic Gaussians centered at each mode. Since $ae_0 + e_1$ is further away from e_2 (distance $\sqrt{a^2 + 2}$) than it is from e_0 (distance $\sqrt{a^2 - 2a + 2}$), it is much less likely under $\mathcal{N}(e_2, \sigma_t)$ than $\mathcal{N}(e_0, \sigma_t)$, leading to a lower weight. This intuition is shown graphically in a 2D projection in Figure 13 (left).

For the detailed proof, we actually want to ensure that p has full support even at $t = 0$ so we add a little bit of noise to it, but choose the covariance such that $z = Ax$ still holds at $t = 0$.

Proof. (Lemma 7.2) Define

$$p_0^0(x) = \frac{1}{2}\mathcal{N}(x; e_0, \tau^2(A^T A)^{-1}) + \frac{1}{2}\mathcal{N}(x; e_2, \tau^2(A^T A)^{-1})$$

$$p_1^0(x) = \frac{1}{2}\mathcal{N}(x; ae_0 + e_1, \tau^2(A^T A)^{-1}) + \frac{1}{2}\mathcal{N}(x; ae_2 + e_3, \tau^2(A^T A)^{-1})$$

$$p_b^0(x) = \mathcal{N}(x; 0, \tau^2(A^T A)^{-1}), \quad \text{where } A := \begin{bmatrix} 1 & -a & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & -a \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

26

so that in the transformed space:

$$(A\sharp p)(z) := p(A^{-1}z), \quad z = Ax$$
$$(A\sharp p_b^0)(z) = \mathcal{N}(z; 0, \tau^2)$$
$$(A\sharp p_0^0)(z) = \frac{1}{2}\mathcal{N}(z; e_0, \tau^2) + \frac{1}{2}\mathcal{N}(z; e_2, \tau^2)$$
$$(A\sharp p_1^0)(z) = \frac{1}{2}\mathcal{N}(z; e_1, \tau^2) + \frac{1}{2}\mathcal{N}(z; e_3, \tau^2).$$

Therefore Lemma 6.1 implies that at time $t = 0$,

$$\hat{\mathcal{C}}[\vec{p}] := \frac{p_0^0(x)p_1^0(x)}{p_b^0(x)} = p_0^0(x|_{(0,2)})p_1^0(x|_{(1,3)}).$$

When we add noise at time $t > 0$ we get:

$$p_i^t(x_t|x_0) := \mathcal{N}(x_t; x_0, \sigma_t^2)$$
$$p_0^t(x) = \frac{1}{2}\mathcal{N}(x_t; e_0, \sigma_t^2 I + \tau^2(A^T A)^{-1}) + \frac{1}{2}\mathcal{N}(x; e_2, \sigma_t^2 I + \tau^2(A^T A)^{-1})$$
$$p_1^t(x) = \frac{1}{2}\mathcal{N}(x_t; ae_0 + e_1, \sigma_t^2 I + \tau^2(A^T A)^{-1}) + \frac{1}{2}\mathcal{N}(x; ae_2 + e_3, \sigma_t^2 I + \tau^2(A^T A)^{-1})$$
$$= \frac{1}{2}A^{-1}(\mathcal{N}(x_t; e_1, \sigma_t^2 A^T A + \tau^2 I) + \mathcal{N}(x; e_3, \sigma_t^2 I A^T A + \tau^2)).$$

We will start by using this counterexample to prove Part 2 of Lemma 7.2, which is the hard part. Note that $\hat{p}^t(x)$ is made up of terms of the following form:

$$\frac{\mathcal{N}(x; \mu_1; C)\mathcal{N}(x; \mu_2, \Sigma)}{\mathcal{N}(x; 0; \Sigma)} = (2\pi)^{-\frac{n}{2}}|\Sigma|^{-\frac{1}{2}}\frac{e^{-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)}e^{-\frac{1}{2}(x-\mu_2)^T\Sigma^{-1}(x-\mu_2)}}{e^{-\frac{1}{2}x^T\Sigma^{-1}x}}$$
$$= (2\pi)^{-\frac{n}{2}}|\Sigma|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1) - \frac{1}{2}(x-\mu_2)^T\Sigma^{-1}(x-\mu_2) + \frac{1}{2}x^T\Sigma^{-1}x\right)$$
$$= (2\pi)^{-\frac{n}{2}}|\Sigma|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}x^T\Sigma^{-1}x + x^T\Sigma^{-1}(\mu_1 + \mu_2) - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 - \frac{1}{2}\mu_2^T\Sigma^{-1}\mu_2\right)$$
$$= C(2\pi)^{-\frac{n}{2}}|\Sigma|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(x-\mu_1-\mu_2)^T\Sigma^{-1}(x-\mu_1-\mu_2)\right)$$
$$= C\mathcal{N}(x; \mu_1 + \mu_2, \Sigma)$$
$$C = \exp\left(-\frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 - \frac{1}{2}\mu_2^T\Sigma^{-1}\mu_2 + \frac{1}{2}(\mu_1 + \mu_2)^T\Sigma^{-1}(\mu_1 + \mu_2)\right)$$
$$= \exp(\mu_1^T\Sigma^{-1}\mu_2)$$

Noting that

$$\tilde{\Sigma}_t := \sigma_t^2 I + \tau^2(A^T A)^{-1} = \sigma_t^2 I + \tau^2\begin{bmatrix} 1+a^2 & a & 0 & 0 \\ a & 1 & 0 & 0 \\ 0 & 0 & 1+a^2 & a \\ 0 & 0 & a & 1 \end{bmatrix}$$

$$\tilde{\Sigma}_t^{-1} = \frac{1}{(a^2+2)\sigma_t^2\tau^2 + \sigma_t^4 + \tau^4}\begin{bmatrix} \sigma_t^2 + \tau^2 & -a\tau^2 & 0 & 0 \\ -a\tau^2 & (a^2+1)\tau^2 + \sigma_t^2 & 0 & 0 \\ 0 & 0 & \sigma_t^2 + \tau^2 & -a\tau^2 \\ 0 & 0 & -a\tau^2 & (a^2+1)\tau^2 + \sigma_t^2 \end{bmatrix},$$

after some algebra we find that $\hat{p}^0(x)$ and $\hat{p}^t(x)$ are both GMMs with same means:

$$\vec{\mu} = \{(1+a)e_0 + e_1, \quad e_0 + ae_2 + e_3, \quad ae_0 + e_2 + e_1, \quad (1+a)e_2 + e_3\},$$

different variances ($\tilde{\Sigma}_0 = \tau^2(A^T A)^{-1}$ and $\tilde{\Sigma}_t$ for all clusters, respectively), and different weights, as follows:

$$\hat{p}^0(x): \quad w^0 = \frac{1}{4}[1,1,1,1]$$

$$\hat{p}^t(x): \quad w^t \propto [M, 1, 1, M], \quad M := \exp\left(\frac{a\sigma_t^2}{(a^2+2)\sigma_t^2\tau^2 + \sigma_t^4 + \tau^4}\right)$$

$$= [1-\varepsilon, \varepsilon, \varepsilon, 1-\varepsilon], \quad \varepsilon := \frac{1}{2}\frac{1}{M+1}.$$

The key idea is that if you compare the weight on the mode at $(1+a)e_0 + e_1$ (which is proportional to M) vs the weight on the mode at $e_1 + ae_2 + e_3$ (proportional to 1) the former is much more likely than the latter as $\sigma_t \to 0$.

The basic idea for lower-bounding the W_2 distance is that w^t has almost no mass on the two of the clusters and so we will need to move a little less than $1/4$ probability over to those clusters. For example we need to move $1/4$ probability onto cluster $e_0 + ae_2 + e_3$ from either $(1+a)e_0 + e_1$ (L2 distance between means is $\sqrt{2a+2}$) or $(1+a)e_2 + e_3$ (L2 distance $\sqrt{2}$). So overall we will have to move a bit less that $1/2$ probability at least $\sqrt{2}$ distance.

To complete the proof we will exploit the *Mixture Wasserstein* distance as an intermediate. We need the following facts from Delon & Desolneux (2020):

$$MW_2(q_0, q_1) := \inf_{\gamma \in \Pi(q_0,q_1) \cap \text{GMM}_{2d}(\infty)} \int \|y_0 - y_1\|^2 d\gamma(y_0, y_1),$$

$$MW_2^2(q_0, q_1) = \min_{c \in \Pi(w_0, w_1)} \sum_{k,l} c_{k,l} W_2^2(q_0^k, q_1^l) \quad \text{(Delon Prop. 4)},$$

$$MW_2(q_0, q_1) \le W_2(q_0, q_1) + 2 \sum_{i=0,1} \sum_{k=1}^{K_i} w_i^k \text{Tr}(\Sigma_i^k) \quad \text{(Delon Prop. 6)},$$

where $\Pi(q_0, q_1)$ denotes the set of all joint distributions with marginals q_0 and q_1, and $\text{GMM}_d(\infty) := \cup_{K \ge 0} \text{GMM}_d(K)$ denotes the set of all finite GMMs. Plus one more handy fact:

$$W_2^2(\mathcal{N}(\mu_x, \Sigma_x), \mathcal{N}(\mu_y, \Sigma_y)) \ge \|\mu_x - \mu_y\|_2^2.$$

$$w^0 = \frac{1}{4}[1,1,1,1]$$

$$w^t = [1-\varepsilon, \varepsilon, \varepsilon, 1-\varepsilon]$$

$$MW_2^2(\hat{p}^0, \hat{p}^t) = \min_{c \in \Pi(w^0, w^t)} \sum_{k,l} c_{k,l} W_2^2(\hat{p}^0[k], \hat{p}^t[l])$$

$$\ge \min_{c \in \Pi(w^0, w^t)} \sum_{k,l} c_{k,l} \|\mu_k - \mu_l\|_2^2$$

$$\ge 2\left(\frac{1}{2} - 2\varepsilon\right) = 1 - 4\varepsilon$$

Above, we noted any $c \in \Pi(w^0, w^t)$ has to move at least $\frac{1}{4} - \varepsilon$ probability each away from indices 1 and 2 in w^0 and onto indices either 0 or 3, and for any of these moves the squared L2 distance is at least 2, i.e. $\|\mu_k - \mu_l\|_2^2 \ge 2$ for $k \in (1,2), l \in (0,3)$. We can use the MW_2 distance to bound the W_2 distance:

$$W_2(\hat{p}^0, \hat{p}^t) \geq MW_2(\hat{p}^0, \hat{p}^t) - 2\sum_{k=0}^{3}(w^0[k]\mathrm{Tr}(\tilde{\Sigma}^0) + w^t[k]\mathrm{Tr}(\tilde{\Sigma}^t))$$
$$\geq MW_2(\hat{p}^0, \hat{p}^t) - 2(\mathrm{Tr}(\tilde{\Sigma}^0) + \mathrm{Tr}(\tilde{\Sigma}^t))$$
$$= (1 - 4\varepsilon)^{\frac{1}{2}} - 2(4\sigma_t^2 + 2\tau^2(4 + 2a^2))$$
$$\geq 1 - 4\varepsilon - 2(4\sigma_t^2 + 2\tau^2(4 + 2a^2)), \quad \forall \varepsilon \leq \frac{1}{4}.$$

Putting everything together, we have

$$W_2(\hat{p}^0, \hat{p}^t) \geq 1 - 4\varepsilon - 2(4\sigma_t^2 + 2\tau^2(4 + 2a^2))$$
$$\varepsilon := \frac{1}{2}\frac{1}{M+1}, \quad M := \exp\left(\frac{a\sigma_t^2}{(a^2+2)\sigma_t^2\tau^2 + \sigma_t^4 + \tau^4}\right).$$

If we set $\sigma_t = \tau$, then

$$W_2(\hat{p}^0, \hat{p}^t) \geq 1 - \frac{2}{\exp(\frac{a}{(a^2+4)\tau^2}) + 1} - (24 + 4a^2)\tau^2$$

Choosing $a = 1$ allows some further simplification:

$$W_2(\hat{p}^0, \hat{p}^t) \geq 1 - \frac{2}{\exp(\frac{1}{5\tau^2}) + 1} - 32\tau^2$$
$$\geq 1 - 33\tau^2, \quad \text{if } \tau^2 < \frac{1}{32}$$
$$\geq 0.5 \quad \text{if } \tau^2 < \frac{1}{66},$$

(in the second-to-last line we used the fact that $\frac{2}{\exp(\frac{1}{5\tau^2})+1} \ll \tau^2$ if $\tau^2 < \frac{1}{32}$, and in the last line we made an arbitrary choice).

We wanted to show that

$$\exists t, t' : \quad W_2(q^t, q^{t'}) \geq \frac{1}{2}\tau^{-1}|t - t'|.$$

Let's use the simple schedule $\sigma_t := t$.

For any $\tau^2 < \frac{1}{66}$, if we pick $t' = 0$ and $t = \tau$, then we have as desired that

$$W_2(\hat{p}^0, \hat{p}^t) \geq 0.5 \equiv 0.5\tau^{-1}|t|.$$

For Part 1 of Lemma 7.2, we need to show that the distributions p_i satisfy: p_i is 1-Lipschitz w.r.t Wasserstein 2-distance:

$$\forall i : \quad W_2(p_i^t, p_i^{t'}) \leq |t - t'|.$$

We can start by proving a Lipschitz upper

$$p(x) := \sum_{k=1}^{K} w_i \mathcal{N}(\mu_k, C_k), \quad x \in \mathbb{R}^n$$

$$p^t(x) := \sum_{k=1}^{K} w_i \mathcal{N}(\mu_k, C_k + \sigma_t^2 I)$$

$$W_2^2(\mathcal{N}(\mu_x, \Sigma_x), \mathcal{N}(\mu_y, \Sigma_y)) := \|\mu_x - \mu_y\|_2^2 + \mathrm{Tr}(\Sigma_x + \Sigma_y - 2(\Sigma_x^{\frac{1}{2}} \Sigma_y \Sigma_x^{\frac{1}{2}})^{\frac{1}{2}})$$

$$:= \|\mu_x - \mu_y\|_2^2 + \|\Sigma_x^{\frac{1}{2}} - \Sigma_y^{\frac{1}{2}}\|_F^2 \quad \text{if } \Sigma_x, \Sigma_y \text{ commute}$$

$$\implies W_2^2(p^{t'}[k], p^t[k]) = \|(C_k + \sigma_t^2 I)^{\frac{1}{2}} - (C_k + \sigma_{t'}^2 I)^{\frac{1}{2}}\|_F^2$$

$$= \|(\Lambda + \sigma_t^2 I)^{\frac{1}{2}} - (\Lambda + \sigma_{t'}^2 I)^{\frac{1}{2}}\|_F^2, \quad \text{where } C_k = U\Lambda U^T \text{ is eigendecomposition}$$

$$\leq \|(\sigma_t - \sigma_{t'})I\|_F^2, \quad \text{(by concavity of square root and } \Lambda \succeq 0)$$

$$= n(\sigma_t - \sigma_{t'})^2$$

$$W_2^2(p^{t'}, p^t) \leq M W_2^2(p^{t'}, p^t)$$

$$:= \min_{c \in \Pi(w,w)} \sum_{k,l} c_{k,l} W_2^2(p^{t'}[k], p^t[l])$$

$$\leq \sum_{k}^{K} W_2^2(p^{t'}[k], p^t[k]), \quad \text{(since } c = I \in \Pi(w,w))$$

$$\leq nK(\sigma_t - \sigma_{t'})^2$$

$$\implies W_2(p^{t'}, p^t) \leq (nK)^{\frac{1}{2}} |\sigma_t - \sigma_{t'}|,$$

showing that p is $(nK)^{\frac{1}{2}}$-Lipschitz w.r.t. W_2 distance. Specializing this to the p_i used in our counterexample where $K = 2$, the Lipschitz constant for each p_i is $\sqrt{2n}$; that is, $\mathcal{O}(1)$ (where $\mathcal{O}$ hides only constants depending on ambient dimension n, and not on τ). $\qquad\square$