RESTORING PLURALITY IN COMPOSITIONAL DIFFUSION BY TRAINING CORRECTION ADAPTERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-to-image diffusion models generate high-fidelity scenes in which many concepts appear together. Given a single prompt such as "a cat and a dog", these methods render both concepts in the same scene as distinct objects that interact naturally. However, scaling the prompt space to include more concepts and their combinations becomes prohibitively expensive. Rather than asking the model to produce a scene from a single prompt with all objects, recent methods evaluate the same model separately on each concept, for example "a cat" and "a dog", then combine the two predictions during inference. Yet inference-time composition remains challenging as models often fail to preserve the intended spatial relationships between objects and collapse the scene into a single hybrid object. In this paper we introduce the term plurality for one aspect of composition, the property that distinct concepts a prompt names appear in the scene as distinct objects. We show that the failure of composition reflects a loss of plurality when two concepts contend for the same image region. This loss is measurable as the plurality term: the per-step residual between the model's prediction under the joint prompt and its combined prediction from the separate prompts. A small fraction of this residual, added back at the right stage of generation, is enough to restore plurality, returning both objects to the scene. To make this correction available without the joint prompt, we train a low-rank adapter on the cross-attention layers of Stable Diffusion XL to predict the residual at every denoising step. We show that the adapter learns a general correction for plurality rather than the appearance of any specific concept, which allows it to transfer to unseen pairs. Overall, our approach takes a step towards obtaining the data efficiency of inference-time composition with the scene coherence of a single joint prompt.

1 INTRODUCTION

Latent diffusion models (LDM) have shown exceptional results for image generation (Sohl-Dickstein et al., 2015; Rombach et al., 2022; Saharia et al., 2022), driven by their ability to seemingly generalise and create completely novel images of known concepts (Biroli et al., 2024). Consequently, these models are being widely adopted and used to generate increasingly complex scenes which often includes multiple interacting objects (Ramesh et al., 2021; Peebles & Xie, 2023; Saharia et al., 2022). However, the number of scenes a text-to-image model can be prompted to generate grows combinatorially as concepts are added to a scene. The most immediate strategy to handle the explosion of possible scene prompts is to gather more data to provide examples of all scenes, requiring increasingly large training datasets to cover the possible combinations (Du & Kaelbling, 2024). This is a sufficient strategy for objects which are often found together physically and are therefore easy to gather sufficiently many examples of in the training data. For example, prompting Stable Diffusion (Rombach et al., 2022) with "a cat and a dog" can produce an image of a physically distinct cat sitting beside a dog. In the absence of enough training examples a different strategy is needed to compose concepts encountered separately in training into novel and coherent scenes. This capacity to combine known pieces into new combinations is termed "compositional generalisation" (Szabó, 2012; Lake, 2019) and is believed to be foundational to human cognition (Lake & Baroni, 2023).

Recent work (Du et al., 2023; Huang et al., 2025) has shown that LDMs with moderate quantities of training data do not generalise compositionally from prompting alone and instead propose methods to explicitly compose concepts at inference time (Liu et al., 2022; Du et al., 2023; Skreta et al.,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

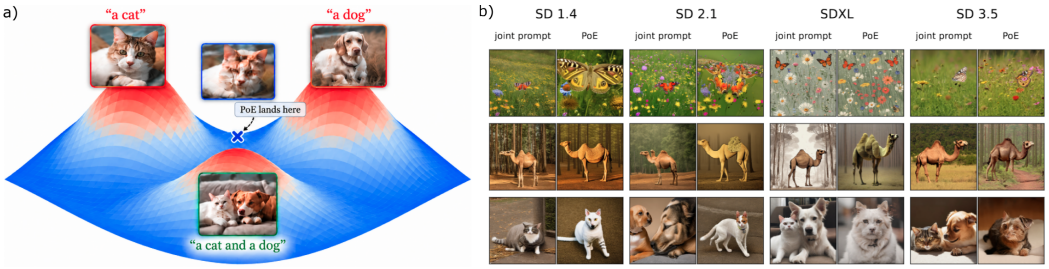

Figure 1: **Product-of-experts composition creates hybrid objects rather than the scene that joint prompts reach.** a) Conditioning the same LDM from equal initial noise with different prompts results in a measurable distance between joint prompting and PoE. PoE and similar compositional diffusion methods combines individual prompts for "a cat" and "a dog" at inference time which can converge to a latent representation between the two individual modes. The distance between this PoE latent representation and the latent representation from the joint prompt encodes the abstract notion of "plurality". b) Plurality and compositional generalisation do not emerge from scale. Each row depicts a prompt pair from (butterfly, flower meadow), (camel, forest) and (cat, dog). Each column set depicts a text-to-image model ordered from oldest to newest. Each image is the model's output for the concept pair, either obtained from the joint prompt (left column) or using PoE (right column). The compositional diffusion rule returns one hybrid animal for a cat and a dog on every model regardless of its size as described on the left.

2025b; Zhang et al., 2025). In these compositional diffusion methods, separate prompts such as "a cat" and "a dog" are passed to the same LDM, and their predictions are combined to guide generation toward a scene containing both concepts. The pretrained model remains unchanged and a new composition requires additional inference rather than additional data and training, mitigating the difficulty in data collection.

Compositional diffusion has been shown to fail when two concepts compete for the same region of the image (Bradley et al., 2025). For example, prompting for a cat and a dog separately implies that both should be the subject of the image, center of frame and in the foreground. Consequently, combining the predictions for competing concepts with a method such as product-of-experts (PoE) (Liu et al., 2022; Zhang et al., 2025) produces samples that score highly under both experts individually. As Figure 1 demonstrates, this is not equivalent to producing a highly likely scene as a whole. The result is that competing concepts tend to get combined into a hybrid object that simultaneously resembles the individual concepts being prompted for (Dutta et al., 2026).

In this work we consider the difference between the LDM output on a single prompt describing the full scene, which we term the *joint prompt*, and the output from compositional diffusion. Our primary concern is with cases where joint prompts succeed on conflicting concept pairs to produce coherent multi-object scenes and compositional diffusion produces a hybrid object. We show that this gap between joint prompting and inference-time composition is measurable in the trajectory of the LDM through latent space. We refer to the lost visual property of having multiple distinct objects as "plurality" and define this concept for the context of compositional diffusion in Definition 1.1.

> **Definition 1.1: Plurality**
>
> The property of a prompt and image pair where each named concept in the prompt is rendered as a distinct object in the generated scene.

We show that plurality can be learned by an adapter network layer with a finite set of examples from conflicting concept pairs. Further, the learned concept of plurality within the latent space generalises to novel concept pairs. Thus, our method is able to induce plurality zero-shot for concept combinations that cannot be found in the training data and that joint prompting is unlikely to generalise to. We make three primary contributions:

- We measure the abstract concept of plurality which is already present in the latent space of a pretrained LDM such as Stable Diffusion XL.

- We show that adding a plurality term to PoE restores composition and that this plurality term is learnable by an adapter network layer, alleviating the need for joint prompting at inference time
- We show that the learned concept of plurality can be generalised to novel concept pairs at inference time, providing PoE with the capacity to generate coherent, multi-object scenes through compositional generalisation of individual objects and the abstract notion of plurality.

2 Background and Related Work

2.1 Latent Diffusion and Classifier-free guidance

A latent diffusion model (Rombach et al., 2022) generates images in the compressed latent space of a variational autoencoder (VAE) (Kingma & Welling, 2013). The forward diffusion process adds Gaussian noise to a latent representation obtained from encoding a real image over successive timesteps. The latent x_t at any timestep t can also be sampled directly from x_0 as:

$$x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I) \tag{1}$$

where $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$, with $\alpha_s = 1 - \beta_s$ and β_s the noise variance added at step s. The sequence of β_s terms parametrize the noise scheduler of the diffusion process, which can have an impact on the performance and stability of LDMs (Hang et al., 2025). During training, a denoising network with parameters θ learns to predict the noise used to produce x_t so that it can be removed, forming the reverse diffusion process. The network's prediction $\epsilon_\theta(x_t, t \mid c)$ given x_t, the time index of the diffusion process t, and a conditioning text prompt c is trained by minimizing the squared error (Ho et al., 2020):

$$\mathcal{L}_{\text{noise}} = \mathbb{E}_{x_0, c, t, \epsilon}\left[\|\epsilon - \epsilon_\theta(x_t, t \mid c)\|_2^2\right]. \tag{2}$$

The noise prediction is related to the distribution of latents at each timestep through its *score*. The score points in the direction where the log density increases most rapidly. The predicted noise approximates the opposite direction, scaled by the noise standard deviation $\sigma_t = \sqrt{1-\bar{\alpha}_t}$ (Song et al., 2021):

$$\epsilon_\theta(x_t, t \mid c) \approx -\sigma_t \nabla_{x_t} \log p_t(x_t \mid c). \tag{3}$$

Here $p_t(x_t \mid c)$ is the distribution of latents at timestep t conditioned on the prompt c. To generate an image, sampling starts from Gaussian noise $x_T \sim \mathcal{N}(0, I)$ and uses the network's predictions to denoise the latent step by step. A decoder turns the resulting x_0 into an image.

Classifier-free guidance (CFG) can strengthen a prompt's influence during sampling (Ho & Salimans, 2021) by training the denoising network to remove noise both with and without text conditioning. In the case of no text conditioning a null-prompt is provided to the model and we denote the model's unconditioned prediction as $\epsilon_\theta(x_t, t)$. During sampling, both predictions can be made at the same latent representation and timestep. Their difference shows how the prompt changes the predicted noise. CFG scales this difference to strengthen the prompt's influence and adds it to the unconditioned prediction:

$$\hat{\epsilon}(x_t, t \mid c) = \epsilon_\theta(x_t, t) + w\left[\epsilon_\theta(x_t, t \mid c) - \epsilon_\theta(x_t, t)\right]. \tag{4}$$

The guidance weight w controls how much the prompt changes the prediction. At $w = 1$, the result is the noise prediction made directly with prompt c. At $w > 1$, the prompt provides a stronger force on the reverse diffusion than if it was only conditioned on, thus enhancing the guidance provided by the prompt. In this paper we use Stable Diffusion XL (Podell et al., 2024) as our latent diffusion model (Rombach et al., 2022) which uses a U-Net as the denoising network architecture (Ronneberger et al., 2015) on a a $4 \times 128 \times 128$ latent representation.

2.2 Compositional Latent Diffusion

Compositional diffusion methods aim to generate an image by containing on multiple prompts c_1 and c_2 separately and without supplying a joint prompt. PoE is one of the most direct methods of achieving this and assumes that the concepts are conditionally independent given an image x (Liu

et al., 2022). This assumption and Bayes' rule allow the joint conditioning distribution to be defined in terms of the individual conditioning distributions:

$$\begin{aligned} p(x \mid c_1, c_2) &\propto p(x)\, p(c_1 \mid x)\, p(c_2 \mid x) \\ &\propto p(x)\, \frac{p(x \mid c_1)}{p(x)}\, \frac{p(x \mid c_2)}{p(x)} \\ &= \frac{p(x \mid c_1)\, p(x \mid c_2)}{p(x)}. \end{aligned} \tag{5}$$

Here x denotes an image. The term $p(x)$ describes the probability density of x under the unconditional distribution over images, while $p(x \mid c_i)$ describes its density when concept c_i is specified. The proportionality omits factors that do not depend on x.

As shown in Equation 3, we aim to compute the score of the data distribution to perform the reverse diffusion process. Taking the log gradient of Equation 5 gives the score for the joint condition:

$$\begin{aligned} \nabla_x \log p(x \mid c_1, c_2) = \nabla_x \log p(x \mid c_1) \\ + \nabla_x \log p(x \mid c_2) - \nabla_x \log p(x). \end{aligned} \tag{6}$$

Given a trained denoising network, this score can be computed directly at each denoising step. The denoising network (a U-Net in the case of Stable Diffusion XL) is evaluated at the same latent representation x_t with prompt c_1, prompt c_2, and the null prompt separately. These predictions can then be combined to approximate the scaled negative score of Equation 6 (Liu et al., 2022):

$$\epsilon_{\mathrm{PoE}}(x_t, t) = \epsilon_\theta(x_t, t \mid c_1) + \epsilon_\theta(x_t, t \mid c_2) - \epsilon_\theta(x_t, t). \tag{7}$$

We omit the reliance on the prompts c_1, c_2 to lighten notation and leave this conditioning implied. The guidance afforded by the individual conditioning prompts can be modulated following a similar reasoning as Equation 4. In the guided PoE rule, each concept has its own weight, while both use the same unconditional prediction. The individual weighted guidance vectors are then added to the unconditioned prediction:

$$\begin{aligned} \hat{\epsilon}_{\mathrm{PoE}}(x_t, t) = \epsilon_\theta(x_t, t) \\ + w_1 \left[\epsilon_\theta(x_t, t \mid c_1) - \epsilon_\theta(x_t, t)\right] \\ + w_2 \left[\epsilon_\theta(x_t, t \mid c_2) - \epsilon_\theta(x_t, t)\right]. \end{aligned} \tag{8}$$

Here w_1 and w_2 are the weights for c_1 and c_2, respectively. At $w_1 = w_2 = 1$, the rule reduces to Equation 7. We set both weights to 7.5 in our experiments.

Recent improvements on PoE keep the product distribution as the target and improve how it is composed and sampled (Dutta et al., 2026; Huang et al., 2026; Wang et al., 2026). Liu et al. (2022) treat a prompt naming two concepts as a conjunction and computes one "expert" prediction per concept. The expert predictions are added directly at each denoising step. Du et al. (2023) show that reverse diffusion on the composed predictions does not sample the product distribution the added scores are meant to aim at, and correct the mismatch with Markov chain sampling. Later methods sharpen the correction, with annealed importance sampling (Zhang et al., 2025) and sequential Monte Carlo correctors (Skreta et al., 2025b;a).

3 The Measured Plurality Term

We define the plurality term r_t as the difference between the model's guided prediction under the joint prompt and its guided product-of-experts prediction. Denoting $\hat{\epsilon}_{\mathrm{Mono}}$ as the joint-prompt prediction and evaluating both model predictions at the same latent representation x_t gives:

$$r_t = \hat{\epsilon}_{\mathrm{Mono}}(x_t, t) - \hat{\epsilon}_{\mathrm{PoE}}(x_t, t). \tag{9}$$

Measuring r_t requires a model which is able to produce a reasonable image conditioned on the joint prompt. This is likely to only be true for common concept combinations and places the burden on data and training, rather than inference, which is counter to the purpose of compositional diffusion. To begin our exploration, we add the measured plurality term to the product-of-experts and control the strength of the correction with λ:

$$\hat{\epsilon}_\lambda(x_t, t) = \hat{\epsilon}_{\mathrm{PoE}}(x_t, t) + \lambda r_t. \tag{10}$$

At $\lambda = 0$, the sampler uses uncorrected composition. At $\lambda = 1$, its prediction at each visited state equals the joint-prompt and we call this the "corrected sampler". This experiment computes r_t at each state visited by the corrected sampler which requires an appropriate prediction under the joint prompt.

Figure 2 shows the effect of sampling with Equation 10 using different correction strengths. This is shown at the top row of generations and we see that the required plurality emerges at an intermediate value of the correct strength, resulting in a type of phase transition. As an ablation we compare against three distortions of the correction 1) a correction term computed using two different concept pair and added back to the original PoE prediction, 2) a correction term computed using the same concept pair from a different generation seed and added back to the original PoE prediction and 3) a correction term computed using the same concept pair and seed but at a diffferent time index and added to the original PoE prediction. Each control matches the measured correction's magnitude at every step. None of the controls produces two animals across the sweep. Consequently, the correction is highly coupled to the state of the image being denoised, but there is a region of the latent space which encapsulates the abstract notion of plurality. This region appears to be distant from the stable point of the original PoE diffusion process.

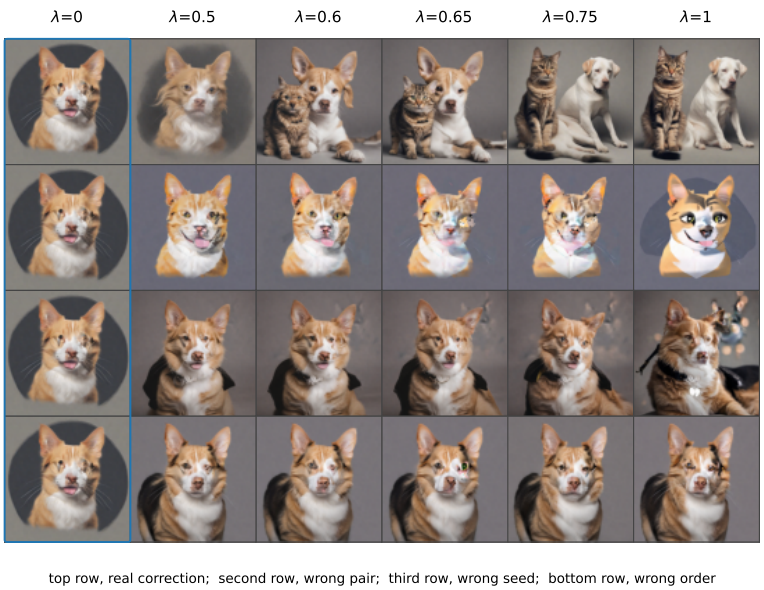

Figure 2: **Only the measured correction composes.** A cat and a dog at seed 9. A column is the strength λ added at every step, up to the full correction at $\lambda = 1$; a row is what was added. The top row is the pair's own plurality term, and beneath it sit three controls, each breaking one thing about it while matching its size and schedule: another pair's term, the same term from a different seed, and the same term with its step order deranged. Nothing is added at $\lambda = 0$, so that column is one sample shown four times.

The strength sweep applies the correction throughout the reverse diffusion process. We now consider whether the correction is needed at every step. To test this we apply the correction term over a subset (or a "stretch") of the denoising time steps. The remaining time steps use the standard PoE denoising rule. Figure 3 compares stretches that begin at the first step with stretches that begin later and continue to the end. Each row uses the same concept pair and starting noise, and its finished image appears at the right-hand edge. Correcting the first ten of fifty steps is enough to finish with a cat beside a dog. Correcting additional steps produces the same outcome. A correction that starts at step twenty runs for thirty steps yet finishes with one blended animal. For this pair and seed, adding more correction later does not recover the outcome lost at the start.

Figure 4 shows a narrower sweep that tests whether this pattern persists across starting noise. We restrict the correction to a ten-step window and move that window across the run in strides of five. The other forty steps use uncorrected PoE composition. A generated image counts as composed when an object detector finds two distinct animals; we manually verified every count. Three of

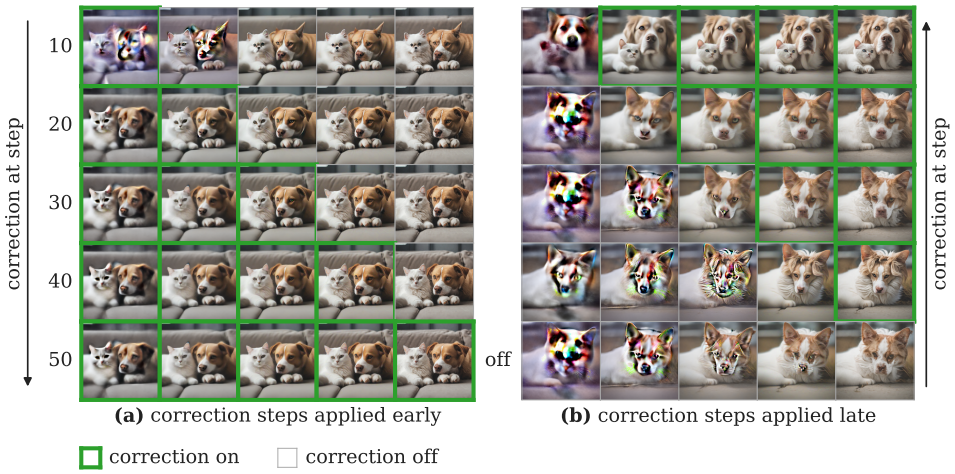

Figure 3: **Ten corrected steps already reach the outcome, and no late start recovers it.** A cat and a dog at seed 12, corrected over one stretch of each run and left to plain composition elsewhere: from the beginning outward in **(a)**, from a later start through to the end in **(b)**. Along any row the five pictures are that run decoded after 10, 20, 30, 40 and 50 steps, a green frame marks a decode taken while the correction was active, and the row marked *off* never corrects. The finished picture of a row is the one at its right-hand edge: down that edge (a) never changes, while (b) changes once, at the row that starts at step twenty.

four seeds compose when the window starts at step 0, and three compose when it starts at step 5. One seed composes when the window starts at step 10 or 15. None composes when it starts at step 20 or later. Taken together, these results show that a correction term exists and that a correction during the early steps of denoising is sufficient to guide product-of-experts composition to an image demonstrating plurality. The experiments above obtain r_t from the joint prompt. We now consider whether this correction can be learned by a LoRA network module (Hu et al., 2022) such that it that can be used without that prompt during sampling.

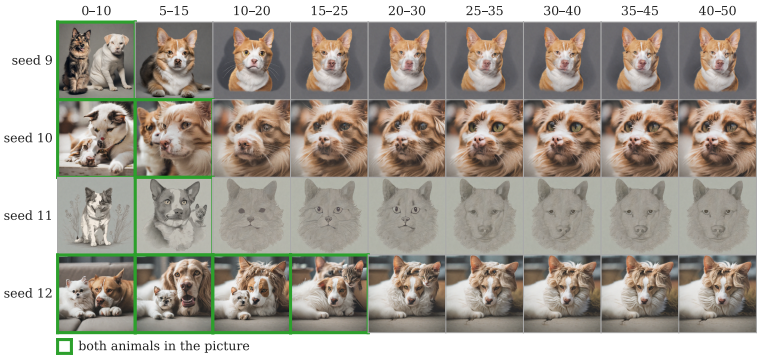

Figure 4: **Composition survives only while the correction window starts early.** A cat and a dog at seeds 9 to 12, as the ten-step window slides across the run. A column is the window's first and last step, and everything outside it runs plain composition, so a row varies only in where the correction is applied. A green frame marks the runs where the detector found both animals, a cat box and a dog box with little overlap. The frames thin out from left to right and stop entirely from step 20 on, with one break: seed 11 composes at 5–15 but not at 0–10.

4 LEARNING THE PLURALITY TERM

We train a LoRA module (Hu et al., 2022) on the query, key, and value projections of every cross-attention of layer of Stable Diffusion XL (Podell et al., 2024) while keeping the model's existing weights frozen. This LoRA module is denoted by $\epsilon_{\theta,\psi}$ with ψ as the LoRA parameters. In these layers, queries are computed from the latent, while keys and values are computed from the prompt tokens. The module is trained to output the plurality correction for a given latent representation. If the module successfully learns the abstract notion of plurality, then it can be included during inference time with the usual PoE denoising step such that a joint prompt is not needed. During training, the joint prompt provides the target. Each network evaluation takes a latent x_t, its timestep t, and one prompt, and returns the predicted correction defined in Equation 9. The adapted composition $\hat{\epsilon}^{\psi}_{\mathrm{PoE}}$ uses separate predictions under c_1, c_2, and the null prompt. It combines them using the same guided product-of-experts rule as the unadapted model with the estimated correction term as shown in Figure 10.

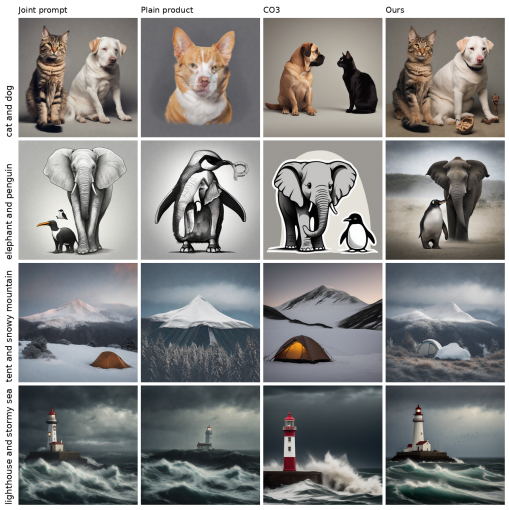

Figure 5: **The adapter repairs the composition without the joint prompt.** Rows are four unseen concept pairs, and columns are the joint prompt, plain product, CO3, and our adapter. CO3 uses different starting noise; the other three columns share it.

More formally, we construct the training pool from concept pairs that Stable Diffusion XL can compose under a joint prompt. A pair that fails under joint prompting cannot provide the desired target. For each candidate pair, we generate one image using plain joint prompting and one using CO3 (Dutta et al., 2026), and retain the pair if either image contains both concepts. This filtering leaves 48 concept pairs which demonstrate plurality from joint prompting. Remaining concept pair where joint prompting cannot even produce an image with plurality are left for evaluation of the generalisability of our LoRA module. Neither concept in a held-out pair appears in any training pair. The starting noise used for evaluation is also excluded from training. For every training pair and seed, we select ten states from an uncorrected product-of-experts trajectory. At each selected (x_t, t), we match the adapted composition to the frozen model's guided prediction under the joint prompt c_{12}:

$$\hat{\epsilon}_{\mathrm{Mono}} = \epsilon_\theta(x_t, t \mid \emptyset) + w\big[\epsilon_\theta(x_t, t \mid c_{12}) - \epsilon_\theta(x_t, t \mid \emptyset)\big]. \tag{11}$$

Only the adapter parameters ψ are updated. At the selected states, the target differs from the unadapted composition by r_t, so matching it trains the adapter to supply that difference. During sampling, the learned correction is the change that the adapter makes to the composed prediction. We denote it by $\hat{r}_\psi = \hat{\epsilon}^{\psi}_{\mathrm{PoE}} - \hat{\epsilon}_{\mathrm{PoE}}$ and apply it in the same form as the measured correction:

$$\begin{aligned} \hat{\epsilon}_\lambda &= \hat{\epsilon}_{\mathrm{PoE}} + \lambda \hat{r}_\psi \\ &= \hat{\epsilon}_{\mathrm{PoE}} + \lambda \left(\hat{\epsilon}^{\psi}_{\mathrm{PoE}} - \hat{\epsilon}_{\mathrm{PoE}} \right). \end{aligned} \tag{12}$$

Appendix A gives the training and sampling details. Appendix B derives the relation between the training target and the prediction used during sampling.

4.1 RESULTS

Figure 5 shows the results of our method compared to joint prompting, PoE and CO3 (which still uses the joint prompt). Importantly, for inference using our LoRA module the joint prompt is never evaluated, and inference still requires only three predictions per denoising step (one per concept and the null prompt where the LoRA is active within these predictions). As expected, joint prompting and CO3 show good performance for common pairs such as the lighthouse and sea which are likely

to be explicitly in the Stable Diffusion XL training set. For uncommon pairs, even joint prompting can fail by including multiple legs on the penguin for example. CO3 correctly depicts a penguin and elephant separately, however completely changes the style to be a cartoon to accommodate this uncommon pairing. Importantly, all examples are generated from the same initial seeds. The second column shows the corresponding product-of-experts samples. In each case, the two concepts are not preserved as distinct entities: the animals fuse, the tent and mountain collapse into one cone-like shape. The lighthouse and sea appears structurally correct but the lighthouse itself is reduced to a distant object in the background. This reproduces the results of Bradley et al. (2025) which show that PoE will combine objects when they compete for the same position in an image and works reasonably well when there is a distinction between which objects should fall into the foreground and background. Finally, the LoRA module trained to restore plurality accurately does so across a range of different concepts. For the cat and dog it almost perfectly recovers the joint prompt (once again without ever having access to this joint prompt, where the consistency is due to the same seed and initial noise sample). For the uncommon pairing of the penguin and elephant it also accurately separate the objects. It is important to note that this pairing is a novel combination that the LoRA was not trained on as there are insufficient examples of a good joint prompt output for this pair (as shown by the failure of the joint prompt here). Thus, our LoRA module successful generalises and produces a structurally sound scene which composes distinct object at inference time.

We also conduct a sensitivity analysis on the one hyper-parameter our method introduces, which is the rank of our LoRA module. Thus, we trained LoRA modules at different ranks to examine whether adapter capacity affects the recovery of two distinct animals. Figure 6 shows this comparison for ranks 8, 16, and 32. Each row uses the same sampled noise and seed. The rank-8 run produces heavily degraded images in which the cat and dog are largely lost in visual noise, while rank-16 and rank-32 runs produce two distinct animals. Thus, recovering plurality does require sufficient rank of the LoRA but beyond sufficient dimensionality performance remains consistent.

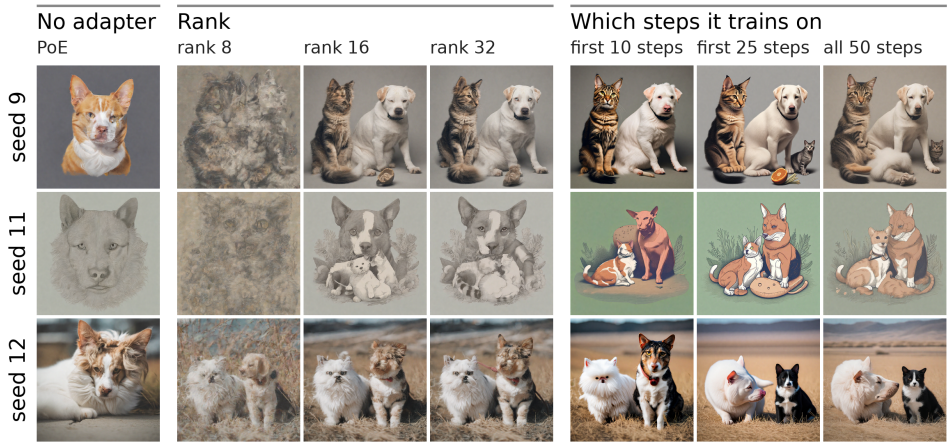

Figure 6: **Effect of adapter settings on composing a cat and a dog** Each row uses one held-out seed and the same starting noise across columns. The left group compares the uncorrected product with adapters of rank 8, 16, and 32. The right group compares correction during the first 10, first 25, or all 50 denoising steps. After the adapter stops, the PoE completes the remaining steps. Therefore, the first column of the right groups shows the adapter acting on the 10 steps followed by 40 steps of uncorrected composition.

Finally, we report two limitation to our approach which should be noted and are demonstrated in Figure 7. Firstly, we consider the effect of applying our LoRA module to a joint prompt. This test whether it can be applied to Stable Diffusion XL as a general property or if it should be used precisely during PoE. We see that the LoRA can change the output from joint prompting and it is not consistently better. For example, the cat and dog image is slightly corrupted and in the tiger and dog pairing it introduces a third animal which itself is a hybrid. However, in the case where the joint prompt fails such as the chessboard and hourglass the LoRA actually improved the image. For easy pairs it seems to have little impact on joint prompting, for example in the picnic basket and watermelon case. Thus, the LoRA module can introduce some instability during joint promting

and should not be used for all image generations. Our LoRA-augmented PoE can also produce nonsensical which typically correspond to intrusion of new animals. For example, when prompted for a dog and a cat it produces a horse and a dog or a person with a dog. Another example introduces a third smaller cat into frame with a dog and a cat, showing that while the LoRA has learned plurality it is not necessarily counting. Finally, if the seed would typically produce a highly stylized image such as a line drawing or artwork the LoRA does not inject plurality. Thus, the LoRA is particularly useful for realistic styles where object positioning and scene structure is a more strict requirement. However, for styles where hybrid animals are more tolerable the LoRA sometimes does not change the final output.

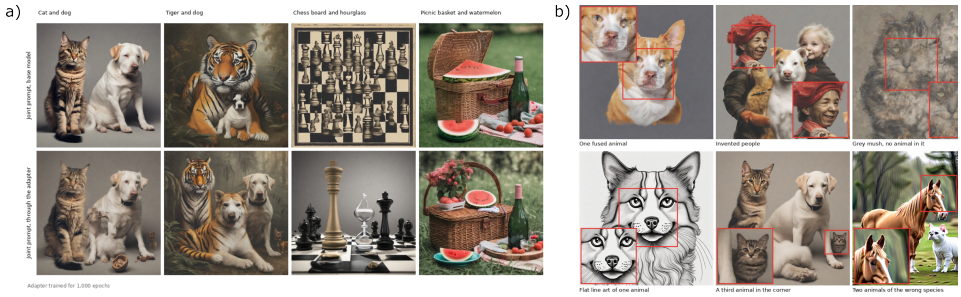

Figure 7: **Limitations of the Plurality LoRA:** a) The adapter can change the render under joint prompting and is inconsistently useful in this case. Each row compares the model conditioned on the joint prompt with and without the LoRA applied from the same noise sample. While the LoRA improves the render for a chessboard and hourglass, it disrupts the render for a cat and a dog, and a tiger and a dog. b) We show five examples of how a corrected cat–dog render can lead to inconsistent or unexpected results. The uncorrected hybrid (top left) is followed by five corrected failures showing intrusions (top middle and bottom right), counting errors (bottom middle), style-dependent loss of plurality (bottom left) or instability (top right).

5 Discussion

In this work we introduce the concept of plurality: the notion that distinct objects in a prompts should become distinct objects in a generated image. We show that PoE and similar compositional diffusion methods are unable to create images with plurality, inside create hybrid objects that combine the individual objects being prompted for. We show that plurality can be restored during early denoising with PoE and train a LoRA module to supply this correction while keeping a base Stable Diffusion XL model frozen. Crucially, the adapter restores composition on concept pairs unseen during training, using only the separate concept prompts and the null token at inference which are needed for PoE. Consequently, by using a finite set of easily obtained examples where joint prompting demonstrates plurality we can train a LoRA to apply the concept more broadly. Thus, our method is able to achieve the coherent image generation of joint prompting without the need for large quantities of data and exhaustive examples of concept pairs.

This work identifies a more general principle: the fact that written text often has implied information which can be learned and encoded by text-to-image generation models. Fundamentally, this is the core result of Section 3 which shows that a joint prompt such as “a cat and a dog” not only encodes the objects which are desire (the cat and the dog independently) but implies a spatial relation between them. When a compositional diffusion method is used to separate the prompt into “a cat” and “a dog” neither individual prompt contains the same implication – neither prompt encodes a restriction on the positioning of the object as, from the perspective of each prompt, there is only one object. The fact that abstract concepts can be encoded and measured in a latent space is not new to diffusion models or this work. Indeed, it is foundational to machine learning, stretching back to seminal work such as Word2Vec (Mikolov et al., 2013). While we have focused on plurality in this work, it is plausible that many more joint prompts can be identified with new examples of implied concepts. Our work here provides a framework for measuring and distilling these concepts into a module which can then be used to augment compositional diffusion in a principled manner. Thus, we take a step towards diffusion which can efficiently compose objects *and abstract concepts* at inference time.

6 AI DECLARATION

In this work we made use of generative AI tools for coding assistance, to polish writing, retrieval of papers and generation of figures. We have not used AI for any other tasks. All AI assisted work was reviewed and manually tested. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Giulio Biroli, Tony Bonnaire, Valentin De Bortoli, and Marc Mézard. Dynamical regimes of diffusion models. *Nature Communications*, 15(1):9957, 2024.

Arwen Bradley, Preetum Nakkiran, David Berthelot, James Thornton, and Joshua M. Susskind. Mechanisms of projective composition of diffusion models. In *Forty-second International Conference on Machine Learning*, 2025.

Yilun Du and Leslie Kaelbling. Compositional generative modeling: A single model is not all you need. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Yilun Du, Conor Durkan, Robin Strudel, Joshua B. Tenenbaum, Sander Dieleman, Rob Fergus, Jascha Sohl-Dickstein, Arnaud Doucet, and Will Grathwohl. Reduce, reuse, recycle: Compositional generation with energy-based diffusion models and MCMC. In *Proceedings of the 40th International Conference on Machine Learning*, 2023. arXiv:2302.11552.

Debottam Dutta, Jianchong Chen, Rajalaxmi Rajagopalan, Yu-Lin Wei, and Romit Roy Choudhury. Steer away from mode collisions: Improving composition in diffusion models. In *International Conference on Learning Representations*, 2026.

Tiankai Hang, Shuyang Gu, Jianmin Bao, Fangyun Wei, Dong Chen, Xin Geng, and Baining Guo. Improved noise schedule for diffusion training. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4796–4806. IEEE, 2025.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, 2020.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. arXiv:2106.09685.

Haozhe Huang, Yudong Xu, Abhijoy Mandal, and Alán Aspuru-Guzik. From global to factor-wise expert composition in discrete diffusion models. *arXiv preprint arXiv:2607.11758*, 2026.

Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2I-CompBench++: An enhanced and comprehensive benchmark for compositional text-to-image generation. *arXiv preprint arXiv:2307.06350*, 2025.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Brenden M Lake. Compositional generalization through meta sequence-to-sequence learning. *Advances in neural information processing systems*, 32, 2019.

Brenden M Lake and Marco Baroni. Human-like systematic generalization through a meta-learning neural network. *Nature*, 623(7985):115–121, 2023.

Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B. Tenenbaum. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision*, pp. 423–439. Springer, 2022.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pp. 8821–8831. Pmlr, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Remi Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo-Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022.

Marta Skreta, Tara Akhound-Sadegh, Viktor Ohanesian, Roberto Bondesan, Alan Aspuru-Guzik, Arnaud Doucet, Rob Brekelmans, Alexander Tong, and Kirill Neklyudov. Feynman-kac correctors in diffusion: Annealing, guidance, and product of experts. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 55906–55949. PMLR, 13–19 Jul 2025a.

Marta Skreta, Lazar Atanackovic, Avishek Joey Bose, Alexander Tong, and Kirill Neklyudov. The superposition of diffusion models using the itô density estimator. In *International Conference on Learning Representations*, 2025b. Spotlight; arXiv:2412.17762.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 2256–2265, 2015.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Zoltán Gendler Szabó. The case for compositionality. 2012.

Zekun Wang, Anant Gupta, Tianyi Zhu, and Christopher J. MacLellan. Test-time compositional generalization in diffusion models via concept discovery. *arXiv preprint arXiv:2605.07078*, 2026.

Yunzhi Zhang, Carson Murtuza-Lanier, Zizhang Li, Yilun Du, and Jiajun Wu. Product of experts for visual generation. *arXiv preprint arXiv:2506.08894*, 2025.

A Implementation details

Every run samples with 50 DDIM steps at guidance weight 7.5, at 1024×1024 resolution, in fp16. The adapter has rank 32 with $\alpha = r$, so its update is added at full scale. It adds 19.8 M trainable parameters to the 2.57 B-parameter frozen base. We train it with AdamW for 50,000 optimizer steps, grouped as 1,000 epochs of 50 steps. The learning rate starts at 10^{-4} and decays to zero on a cosine schedule, and gradients are clipped at norm 1.0.

Figure 8 shows how these runs are trained. Frozen SDXL first supplies the joint-prompt target on states cached from a plain product-of-experts trajectory. At each selected state, the adapted network evaluates the two concept prompts and the null token separately; their composed prediction is compared with that target. The loss updates only the LoRA weights. The diagram separates the one-time target cache from the three adapted predictions and the loss used to update the LoRA weights.

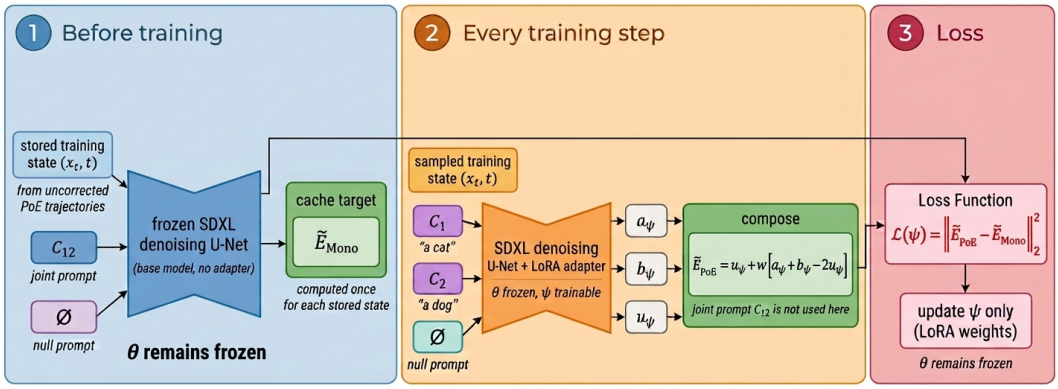

Figure 8: **Training the compositional adapter.** A cached frozen joint-prompt target supervises the adapted product-of-experts prediction; only the LoRA weights are updated. The frozen network supplies the cached target; the adapted network produces the composition matched to it.

B Training and inference derivation

The base network has frozen parameters θ; only the LoRA parameters ψ change during training. We use $c_1 =$ "a cat" and $c_2 =$ "a dog" as an example, write $c_{12} =$ "a cat and a dog" for the joint prompt, and denote the null token by $\emptyset$. Every prediction below is evaluated at the same noisy latent x_t and timestep t.

B.1 Composing the adapted predictions

The adapter never receives c_{12}. It receives c_1, c_2, and $\emptyset$ in three separate forward passes, producing

$$a_\psi = \epsilon_{\theta,\psi}(x_t, t \mid c_1), \quad b_\psi = \epsilon_{\theta,\psi}(x_t, t \mid c_2), \quad u_\psi = \epsilon_{\theta,\psi}(x_t, t \mid \emptyset). \tag{13}$$

With classifier-free guidance weight w, each concept prediction is guided against the same null prediction: $g_1 = u_\psi + w(a_\psi - u_\psi)$ and $g_2 = u_\psi + w(b_\psi - u_\psi)$. Adding the two guided predictions counts the shared null prediction twice, so one copy is subtracted:

$$\begin{aligned} \hat{\epsilon}^{\psi}_{\text{PoE}} &= g_1 + g_2 - u_\psi \\ &= u_\psi + w(a_\psi - u_\psi) + w(b_\psi - u_\psi) \\ &= u_\psi + w(a_\psi + b_\psi - 2u_\psi). \end{aligned} \tag{14}$$

At $w = 1$, this is $a_\psi + b_\psi - u_\psi$; the runs in this paper use $w = 7.5$.

B.2 FROM THE TRAINING TARGET TO INFERENCE

For comparison, write the frozen network's three separate predictions as $a_0 = \epsilon_\theta(x_t, t \mid c_1)$, $b_0 = \epsilon_\theta(x_t, t \mid c_2)$, and $u_0 = \epsilon_\theta(x_t, t \mid \emptyset)$. Its product-of-experts prediction is

$$\hat{\epsilon}_{\mathrm{PoE}} = u_0 + w(a_0 + b_0 - 2u_0). \tag{15}$$

The frozen network also evaluates the joint prompt during training. Writing $j_0 = \epsilon_\theta(x_t, t \mid c_{12})$, its guided joint-prompt target is

$$\hat{\epsilon}_{\mathrm{Mono}} = u_0 + w(j_0 - u_0). \tag{16}$$

The plurality term is the difference between these two frozen predictions at the same state:

$$r_t = \hat{\epsilon}_{\mathrm{Mono}} - \hat{\epsilon}_{\mathrm{PoE}}, \qquad \hat{\epsilon}_{\mathrm{Mono}} = \hat{\epsilon}_{\mathrm{PoE}} + r_t. \tag{17}$$

Training matches the adapted composition to the frozen joint-prompt target. Consequently, $\hat{\epsilon}^{\psi}_{\mathrm{PoE}} \approx \hat{\epsilon}_{\mathrm{PoE}} + r_t$ on the training states. Defining the change made by the adapter as

$$\hat{r}_\psi = \hat{\epsilon}^{\psi}_{\mathrm{PoE}} - \hat{\epsilon}_{\mathrm{PoE}}, \qquad \hat{\epsilon}^{\psi}_{\mathrm{PoE}} = \hat{\epsilon}_{\mathrm{PoE}} + \hat{r}_\psi, \tag{18}$$

shows what training asks it to supply: $\hat{r}_\psi \approx r_t$. The equality in the second expression is exact by definition; the approximation depends on how well the adapter matches the target. The early denoising steps are where this agreement matters most for composition.

At inference, sampling starts from $x_T \sim \mathcal{N}(0, I)$ and evaluates only c_1, c_2, and $\emptyset$ at each step. Their adapted predictions give $\hat{\epsilon}^{\psi}_{\mathrm{PoE}}$ directly, without evaluating c_{12}. The joint-prompt target supplies supervision during training, while sampling retains the same three prompt evaluations per step as ordinary product-of-experts composition.