Coarse-to-Fine Compositional Diffusion for Long-Horizon Planning

Byoungwoo Park[1,2] **Utkarsh A. Mishra**[2] **Jaemoo Choi**[2] **Juho Lee**[1] **Yongxin Chen**[2]
[1]KAIST [2]Georgia Institute of Technology

Abstract

Diffusion models provide strong priors for generating structured data, but many tasks require outputs beyond the scale on which these models are typically trained. Compositional generation addresses this by composing overlapping local plans from a pretrained short-horizon prior into a long-horizon output. However, standard composition primarily enforces agreement between neighboring local plans, yielding local consistency without directly specifying the global structure of the full composition. As a result, locally compatible plans may still form an implausible route, task sequence, or temporal evolution. Existing methods improve global coherence by repeatedly propagating local consistency signals or by adding inference-time optimization, but these procedures become expensive as the number or dimensionality of local plans increases. We propose Coarse-to-Fine Compositional Diffusion (CoFi), an inference-time sampler that separates global structure formation from local detail refinement. CoFi first aligns local denoised estimates around a shared coarse structure, producing a global scaffold that captures the long-range task-level arrangement. It then diffuses this scaffold to an intermediate noise level and denoises it with the same pretrained local prior, restoring local fine structure while preserving the scaffold-induced global coherence. Across long-horizon robotic planning, panoramic image generation, and long video generation, CoFi not only improves both global coherence and local sample quality over prior compositional baselines, but also requires 2–8$\times$ fewer denoiser evaluations.

Project page: https://CoFi-diffusion.github.io/

1 Introduction

Diffusion models provide strong priors for generating structured data (Ho et al., 2020; Song et al., 2021), but many tasks require outputs beyond the scale on which these models are typically trained. For example, a robot may need to reach a distant goal by chaining short feasible motions into a long plan (Luo et al., 2026). Directly training a model for such long-horizon outputs is often impractical, since long-horizon data are scarce and the modeling cost grows quickly with sequence length or spatial extent (Du and Kaelbling, 2024). Compositional generation offers a practical alternative. Instead of training a long-horizon model, it generates overlapping local plans from a pretrained short-horizon prior and composes them into a single long-horizon plan (Zhang et al., 2023; Mishra et al., 2023) as shown in Figure 1. The same pretrained local prior can therefore be reused beyond its original generation scale, and the local plans can be generated in parallel.

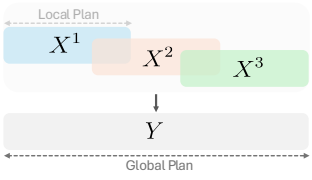

Figure 1: The global plan Y is obtained by composing overlapping local plans X^i.

A typical approach to composition is to encourage neighboring local plans to agree at their shared overlaps during the reverse diffusion process (Zhang et al., 2026). This local coupling reduces boundary mismatch and produces smooth transitions between adjacent local plans, which we refer to

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

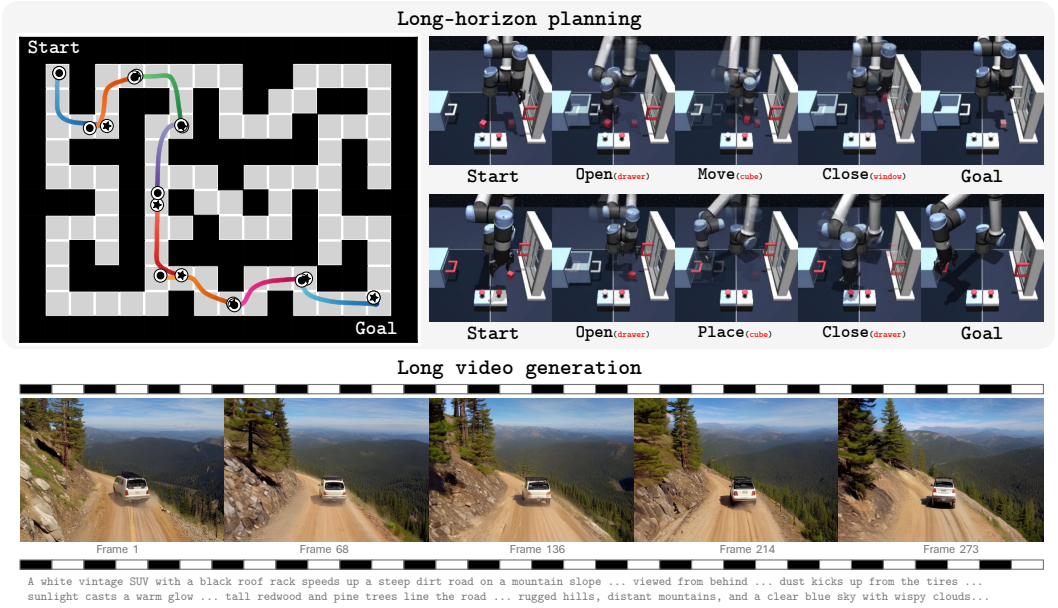

Figure 2: **Scaling short-horizon priors by composition.** A long structured global plan is built by composing overlapping local plans from a pretrained short-horizon prior. The goal is not only to enforce *local consistency* (*e.g.*, smooth transitions between adjacent local plans), but also to preserve *global coherence* (*e.g.*, a plausible route, task sequence, or subject identity over the full horizon). We aim for a simple and effective inference-time sampler that satisfies both desired properties.

as *local consistency*. However, *local consistency* does not guarantee *global coherence*: A composition can match every overlap and still lack a plausible task-level structure across the full horizon. For example, adjacent local plans may connect smoothly at their boundary yet follow incompatible routes that never reach the goal. In long videos, consecutive clips may align at their overlap while the subject's identity gradually drifts over time. A compositional sampler must therefore satisfy a stronger requirement than boundary matching alone: adjacent local plans should remain mutually compatible, while their composition should induce a globally plausible long-horizon structure.

Existing methods attempt to improve the global coherence by making the local consistency signal propagate more broadly across the local plans. Some methods repeat the overlap coupling multiple times within each denoising step, so that information can be passed from one segment to the next over longer distances (Mishra et al., 2026). Other methods introduce inference-time optimization, often using a task-level objective or reward function, to directly adjust the local plans toward a more globally coherent composition (Lee et al., 2023; Kim et al., 2025). While both strategies can improve *global coherence*, they become expensive as the number of the local plans or the dimensionality of each segment increases. This limitation motivates the following design question:

Can we design a simple and effective compositional diffusion sampler that ensures both local consistency and global coherence without expensive inner iterations or optimization?

In this paper, we propose a simpler solution: instead of repeatedly propagating local consistency signals through neighboring overlaps, we first align the local denoised estimates around a *shared coarse structure*. This stage exposes all local plans to a common global signal, promoting *global coherence*, but it can also suppress plan-specific fine details. We therefore treat the resulting coarse structure as a scaffold rather than as the final sample. To restore the suppressed local structure, we diffuse the scaffold forward to an intermediate noise level and denoise it again with the same pretrained local prior, analogous to (Meng et al., 2022). Intuitively, the first stage determines the task-level arrangement of the composition, while the second stage restores local fine structure.

We call this procedure *Coarse-to-Fine Compositional Diffusion* (**CoFi**), an inference-time composition sampler for long-horizon generation with pretrained short-horizon priors. This two-stage design directly constructs a global scaffold, while adding only a partial second denoising pass for refinement. We evaluate CoFi on long-horizon robotic planning, panoramic image generation, and long video generation. Across these domains, CoFi improves both global coherence and local sample quality over prior compositional baselines, *while requiring* 2–8$\times$ *fewer denoiser evaluations* than baseline.

2 Preliminaries

Compositional generation We consider the problem of constructing a long-horizon structured *global plan* $Y \in \mathbb{R}^d$ from a collection of short-horizon *local plans* $X \in \mathbb{R}^m$. Direct samples from the full long-horizon distribution $p(Y)$ are scarce, whereas short local behaviors are available and can be modeled reliably. We therefore assume access to a pretrained generative prior over local plans:

$$X^i = [x^{2i-1}, x^{2i}, x^{2i+1}], \quad i = 1, \dots, N, \tag{1}$$

where N is the number of plans. Here, the long-horizon structure arises through shared boundary factors between consecutive plans. Let $\mathtt{L}$ and $\mathtt{R}$ denote linear selectors that extract the *left* and *right* boundary factors of X^i, respectively:

$$\mathtt{L}(X^i) = x^{2i-1}, \quad \mathtt{R}(X^i) = x^{2i+1}, \quad \forall i = 1, \dots, N-1 \tag{2}$$

A valid composition requires neighboring local plans to agree on their shared *overlapping* boundary,

$$\mathtt{L}(X^{i+1}) = \mathtt{R}(X^i) = x^{2i+1}, \quad \forall i = 1, \dots, N-1 \tag{3}$$

This induces a chain over sequence of *the unique factors*, resulting the long-horizon global plan:

$$Y = \mathtt{Compose}(X^1, \dots, X^N) = [x^1, x^2, \dots, x^{2N+1}]. \tag{4}$$

More generally, $\mathtt{Compose}(\cdot) : \mathbb{R}^{m \times N} \to \mathbb{R}^d$ merges the overlapping local plans into a single global objective, either by using agreed boundary factors (Luo et al., 2026) or by averaging overlapped regions (Zhang et al., 2023; Mishra et al., 2023). This merge step handles *local consistency*, but it does not determine the global structure of the composed plan. Even when every adjacent pair agrees at the overlap, the full plan Y can still drift over long horizons in style, identity, or task-level structure. We refer to this nonlocal requirement as *global coherence*.

Local diffusion prior In this paper, we focus on a diffusion prior for the local plan. Let $X_0 \in \mathbb{R}^m$ denote a local plan drawn from $p_{\mathtt{data}}$. Following (Ho et al., 2020), the forward process is given by

$$q(X_t \mid X_0) = \mathcal{N}\big(X_t; \sqrt{\bar{\alpha}_t} X_0, (1 - \bar{\alpha}_t) I\big), \tag{5}$$

where $\bar{\alpha}_t := \prod_{s=1}^{t} (1 - \beta_s)$, with $\beta_s \in (0, 1)$. In other words, $X_t = \sqrt{\bar{\alpha}_t} X_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon$ so that X_t can be viewed as progressively interpolating between the clean local plan X_0 and Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$. This places diffusion models within the broader view of stochastic interpolants and flow-based generative dynamics (Albergo et al., 2025; Lipman et al., 2023). The reverse process is approximated with a neural network $\epsilon_\theta(X_t, t) \approx -\sqrt{1 - \bar{\alpha}_t} \nabla \log p_t(X_t)$ and sampling proceeds through a reverse transition parameterized by DDIM framework (Song et al., 2021):

$$X_{t-1} = \mathtt{DDIM}_\theta(\hat{X}_{0|t}, t) := \sqrt{\bar{\alpha}_{t-1}} \hat{X}_{0|t} + \sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2}\, \epsilon_\theta(X_t, t) + \sigma_t z, \quad z \sim \mathcal{N}(0, I_m), \tag{6}$$

$$\hat{X}_{0|t} := \tfrac{1}{\sqrt{\bar{\alpha}_t}} \left(X_t + (1 - \bar{\alpha}_t) \nabla \log p_t(X_t)\right) \approx \tfrac{1}{\sqrt{\bar{\alpha}_t}} \left(X_t - \sqrt{1 - \bar{\alpha}_t}\, \epsilon_\theta(X_t, t)\right) \tag{7}$$

where $\sigma_t = \eta \sqrt{\frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t}} \sqrt{1 - \frac{\bar{\alpha}_t}{\bar{\alpha}_{t-1}}}$. This formulation is useful for compositional generation, since the reverse process admits direct inference-time modification through additional guidance terms.

It is worth noting that, if we apply the reverse transition in (6) independently to each local plan, the induced model factorizes as $\prod_{i=1}^{N} p_t(X_t^i)$, and ignores that the local plans should be different views of the same global plan Y_t. The resulting samples can be plausible as individual local plans, but still disagree at their overlaps or fail to form a coherent long-horizon object after composition.

Guided Compositional Diffusion The ideal reverse process would sample from a locally consistent and yet globally coherent distribution over global plan Y_t, rather than from independent local priors $\prod_{i=1}^{N} p_t(X_t^i)$. In principle, one would like to guide the reverse process with an objective that captures both desired properties. A natural way to couple the local reverse processes is to tilt by a reward:

$$p(Y_t) \propto e^{r_t(Y_t)} \textstyle\prod_{i=1}^{N} p(X_t^i), \tag{8}$$

where composition reward r_t penalizes invalid compositions. Ideally, this reward would capture both *local consistency* among neighboring local plans X_t^i and *global coherence* of the composed global plan Y_t. However, designing such a reward is difficult because these requirements are qualitatively different: *local consistency* is pairwise and overlap-based, whereas *global coherence* is nonlocal and depends on the structure of the entire composition (*i.e.*, plausible task-level plan).

This makes the direct design of a tractable (and differentiable) compositional reward r_t for both desired properties nontrivial, especially in high-dimensional plan spaces. For this reason, existing guided compositional samplers usually replace r_t with local surrogates $r_t(Y_t) := \sum_{i=1}^{N} \hat{r}_t\left(X_t^i, X_t^{i+1}\right)$. For example, a composition technique (Zhang et al., 2026) imposes boundary agreement on the predicted clean plans $\hat{X}_{0|t}$ in (7). In our notation, this corresponds to an overlap reward:

$$\hat{r}_t\left(X_t^i, X_t^{i+1}\right) = -\left\|\mathtt{R}(\hat{X}_{0|t}^i) - \mathtt{L}(\hat{X}_{0|t}^{i+1})\right\|^2. \tag{9}$$

Additionally, the Bethe-style composition (Zhang et al., 2023; Mishra et al., 2023) can be viewed as another local surrogate, where the gradient of the surrogate reward $\hat{r}_t$ is approximated using the two adjacent score:

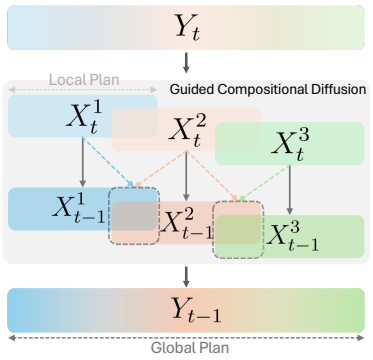

Figure 3: Local guidance reduces boundary mismatch but leaving the global coherence under-specified.

$$\nabla \hat{r}_t\left(X_t^i, X_t^{i+1}\right) \approx -\tfrac{1}{2}\left[\nabla \log p_t\left(\mathtt{R}(X_t^i)|X_t^i\right) + \nabla \log p_t\left(\mathtt{L}(X_t^{i+1})|X_t^{i+1}\right)\right], \tag{10}$$

These local surrogates provide a tractable guidance signal for the gradient of the global composition reward $r_t(Y_t)$ in (8). In practice, the local reverse transition in (6) is guided by the surrogate reward:

$$X_{t-1}^i = \mathtt{DDIM}_\theta(\hat{X}_{0|t}^i, t) + \sigma_t \nabla \hat{r}_t(X_t^i, X_t^{i+1}), \tag{11}$$

where $\mathtt{DDIM}(\cdot,\cdot)$ denotes the unguided reverse transition in (6). This guidance is cheap and explicit, but it only communicates information through adjacent overlaps. It can reduce boundary mismatch, but it does not directly determine the coarse global structure of Y_t.

3 Coarse-to-Fine Compositional Diffusion

In this section, we introduce CoFi, *a simple and effective inference-time diffusion framework for compositional generation* with pretrained local priors. CoFi follows a coarse-to-fine design: it first constructs *a coarse global scaffold* that captures the long-range arrangement, and then refines this scaffold with the same pretrained local diffusion prior to recover *fine local details*.

Most tractable guidance objectives (Mishra et al., 2023; Zhang et al., 2026) introduced in Section 2 act through local consistency as illustrated in Figure 3, with canonical choices such as the Bethe approximation in (10). They encourage neighboring local plans to agree, but they do not directly specify whether the full composition has a plausible task-level structure. A single local guidance step in (11) only couples adjacent overlaps, leaving distant local plans weakly coordinated. Recent methods address this limitation by repeating the local coupling inside the sampling process. For example, CDGS (Mishra et al., 2026) improves global coherence through inner iterations. These steps can be viewed as a form of iterative message passing: information first moves between neighboring local plans, and repeated updates gradually expand the effective communication range across the chain. This is effective, but the cost grows with the number of local plans and becomes especially large for high-dimensional compositions such as visual planning.

In this paper, we aim to develop a simple procedure that remains computationally efficient while satisfying both local consistency and global coherence. To do so, we replace repeated local message passing with a shared scaffold over the local denoised estimates. This gives all local plans access to a common coarse signal, forming a global scaffold. Because the global arrangement has already been outlined, refinement only needs to restore local fine structure rather than propagate global constraints again. We therefore diffuse the scaffold forward to an intermediate noise level $t^\star \in \{1, 2, \ldots, T\}$ and denoise it again using the local guided step in (11). This diffusion refinement step follows the same principle as (Meng et al., 2022): keep the coarse structure, then let the diffusion prior restore details.

Coarse global plan construction. We first define the scaffold alignment update that produces the coarse global scaffold. For each denoising step t, the clean estimate for each local plan $\hat{X}_{0|t}^i$ may be individually plausible, yet still fail to induce a coherent global composition. We therefore seek a coarse estimate that retains the shared global structure while attenuating plan-specific variation.

Algorithm 1 Coarse Scaffold Construction

Input: local denoiser ϵ_θ, local guidance reward $\hat{r}_t$

Initialize local plans $X_T^1, \dots, X_T^N \sim \mathcal{N}(0, I)$
for $t = T, \cdots, 1$ **do**
  for $i = 1, \cdots, N$ **do in parallel**
    Compute Tweedie estimate $\hat{X}^i_{0|t}$ using (7)
    Estimate $\bar{X}^i_{0|t}$ using (13)
    Update X^i_{t-1} using guided step in (14)
  end for
end for
Compose coarse global plan Y_0^{c} using (15).

Output: coarse global plan Y_0^{c}

Algorithm 2 Structure-Preserving Refinement

Input: coarse global plan Y_0^{c}, refinement timestep $t^\star$, local denoiser ϵ_θ, local guidance reward $\hat{r}_t$

Get noised scaffold $Y^{\mathrm{r}}_{t^\star}$ by noising Y_0^{c} using (16)
Let $\{X^{i,\mathrm{r}}_{t^\star}\}_{i=1}^N$ be the local views induced by $Y^{\mathrm{r}}_{t^\star}$
for $t = t^\star, \cdots, 1$ **do**
  for $i = 1, \cdots, N$ **do in parallel**
    Compute Tweedie estimate $\hat{X}^{i,\mathrm{r}}_{0|t}$ using (7)
    Update $X^{i,\mathrm{r}}_{t-1}$ using the guided step in (11)
  end for
end for
Compose refined global plan Y_0^{r} using (15).

Output: final global plan Y_0^{r}

To do so, the coarse estimate should satisfy two competing requirements: for each local plan X^i, it should remain close to the denoised estimate $\hat{X}^i_{0|t}$ so that local plausibility is preserved, while at the same time admitting a decomposition into a shared global component and plan-specific residuals. To formalize this trade-off, we define a simplified scaffold alignment objective for each local plan:

$$\min_{Z^1,\cdots,Z_N} \sum_{i=1}^{N} \left[\|Z^i - \hat{X}^i_{0|t}\|^2 + \gamma_t \|Z^i - \bar{Z}\|^2 \right], \qquad \bar{Z} := \frac{1}{N}\textstyle\sum_{i=1}^N Z^i \tag{12}$$

where γ_t controls the trade-off between local fidelity and shared component. The first term preserves local plausibility, whereas the second promotes a shared global scaffold across local plans. Since (12) is a quadratic problem, its minimizer admits the closed-form solution below, which interpolates each local estimate toward the shared scaffold. Let $Z^{i,\star} := \bar{X}^i_{0|t}$ be the minimizer of (12). Then:

$$\bar{X}^i_{0|t} = (1 - \lambda_t)\bar{X}_{0|t} + \lambda_t \hat{X}^i_{0|t}, \quad \bar{X}_{0|t} := \frac{1}{N}\textstyle\sum_{i=1}^N \hat{X}^i_{0|t} \tag{13}$$

where $\lambda_t := \frac{1}{1+\gamma_t} \in [0, 1]$. If $\lambda_t = 1$ (*i.e.*, $\gamma_t = 0$ in (12)), the scaffold term has no effect and $\bar{X}^i_{0|t} = \hat{X}^i_{0|t}$. If $\lambda_t = 0$ (*i.e.*, $\gamma_t \to \infty$ in (12)), all local estimates collapse to the shared scaffold $\bar{X}_{0|t}$. In other words, each optimal coarse estimate $\bar{X}^i_{0|t}$ is obtained by shrinking the local denoised estimate $\hat{X}^i_{0|t}$ toward the shared scaffold mean. The shared global component $\bar{X}_{0|t}$ is preserved, while the plan-specific residual $\hat{X}^i_{0|t} - \bar{X}_{0|t}$ is attenuated by a factor of λ_t. We then replace the local denoised estimate $\hat{X}^i_{0|t}$ in (11) with scaffold alignment $\bar{X}^i_{0|t}$ for global coherency:

$$X^i_{t-1} = \underbrace{\texttt{DDIM}_\theta(\bar{X}^i_{0|t}, t)}_{\texttt{global}} + \underbrace{\sigma_t \nabla \hat{r}_t(X^i_t, X^{i+1}_t)}_{\texttt{local}}, \tag{14}$$

As a result, the reverse process in (14) is biased toward a globally aligned but locally coarse scaffold. This is precisely the role of the coarse global plan: it suppresses incompatible local variation while preserving the long-range structure shared across local plans, leaving fine local detail to the subsequent refinement stage. Repeating this update from $t = T$ to $t = 0$ gives a coarse-structured global plan:

$$Y_0^{\mathrm{c}} = \texttt{Compose}(X_0^1, \dots, X_0^N). \tag{15}$$

Here, the superscript c denotes the coarse stage. The resulting coarse global plan Y_0^{c} captures the global arrangement of the composition, but may lack fine local detail because the scaffold construction stage softens local fine details by the timestep-dependent factor λ_t.

Structure-preserving diffusion refinement. Starting from the coarse global plan Y_0^{c}, we re-inject noise at an intermediate timestep $t^\star$ and run a second reverse diffusion pass. The purpose of this stage is to recover locally plausible detail under the pretrained diffusion prior while preserving the global arrangement induced by the coarse scaffold. More concretely, the coarse global plan determines *where* the global structure should lie, whereas the second reverse diffusion pass determines *how* the

Table 1: **Quantitative results on compositional planning.** We report OGBench (Park et al., 2025) success rates over 100 evaluation episodes averaged across 5 seeds. Best results are highlighted.

Env	Size	GCBC	GCIVL	GCIQL	HIQL	GSC	CD	CDGS	**CoFi**
`Pointmaze Stitch`	`Medium`	23±18	70±14	21±9	74±6	100±0	100±0	100±0	100±0
	`Large`	7±5	12±6	31±2	13±6	100±0	100±0	100±0	100±0
	`Giant`	0±0	0±0	0±0	0±0	29±3	68±3	82±4	96±2
`Antmaze Stitch`	`Medium`	45±11	44±6	29±6	94±1	96±2	97±2	97±1	97±2
	`Large`	3±3	18±2	7±2	67±5	66±2	86±2	88±2	88±2
	`Giant`	0±0	0±0	0±0	21±2	20±1	65±3	84±3	85±3
`Scene Play`	-	5±1	42±4	51±4	38±3	8±2	13±1	51±2	63±3

final composition should appear at the local level. Formally, after the coarse stage produces the global plan Y_0^{c} in (15), we diffuse it forward to an intermediate timestep $t^\star \in \{1, 2, \ldots, T\}$:

$$Y_{t^\star}^{\mathrm{r}} = \sqrt{\bar{\alpha}_{t^\star}} Y_0^{\mathrm{c}} + \sqrt{1 - \bar{\alpha}_{t^\star}} \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, I_d). \tag{16}$$

The forward noising step in (16) is applied to the coarse global scaffold Y_0^{c}, rather than separately to each local plan X_0^i. If the local plans were noised independently, each plan would receive its own perturbation, which could break the alignment produced by the scaffold construction stage in (14). Noising the global scaffold with a shared noise ε in (16) avoids this issue: the local plans are induced from the same prior $Y_{t^\star}^{\mathrm{r}}$ and start as compatible views of a shared global plan. It restores local fine structure (Meng et al., 2022) while keeping the reverse process anchored to the global plan in (15).

After this structure-preserving noising step, we continue the reverse process locally. Let $X_{t^\star}^{i,\mathrm{r}}$ denote the i-th local plan induced by the refined global plan $Y_{t^\star}^{\mathrm{r}}$. Starting from $t^\star$, each local plan follows the guided composition in (11). We summarize the two-stage sampling procedure in Algorithms 1 & 2.

4 Experiments

We evaluate CoFi on three compositional generation domains: (1) long-horizon robotic planning, (2) panoramic image generation, and (3) long video generation. We adopt the experimental protocol of Mishra et al. (2026) and use the same pretrained diffusion models, datasets, and evaluation metrics.

- *Robotic planning.* We evaluate on OGBench (Park et al., 2025), which includes stitch and play datasets for `PointMaze`, `AntMaze`, and `Scene` manipulation tasks. The stitch datasets contain short maze trajectories, while the play datasets contain unstructured interaction data. Neither directly solves the evaluation tasks, so successful planning requires composing short fragments into a longer plan. We report success rates over 100 episodes averaged across 5 seeds. For maze tasks, we follow Luo et al. (2026) and use distance-based replanning.
- *Panoramic images.* We use Stable Diffusion 2.0 (Rombach et al., 2022) as the local image prior. Each panorama is generated by composing 9 overlapping 512×512 patches into a 512×4608 image with 50% overlap. We evaluate 14 prompts and report Intra-LPIPS (Zhang et al., 2018), Intra-Style loss (Gatys et al., 2016), and CLIP score (Radford et al., 2021).
- *Long videos.* We use CogVideoX-2B (Yang et al., 2025) as the local video prior. The base model natively generates short clips of 50 frames, and we extend it to 273 frames at 720p by composing 9 temporal chunks with 50% overlap. We evaluate 8 prompts using VBench (Huang et al., 2024), reporting subject consistency, temporal flickering, aesthetic quality, and prompt alignment.

Baselines. We compare with both non-compositional and compositional baselines. For robotic planning, we include goal-conditioned methods, GCBC (Lynch et al., 2020; Ghosh et al., 2021), GCIVL, GCIQL (Kostrikov et al., 2021), and HIQL (Park et al., 2023), as well as compositional methods, GSC (Mishra et al., 2023), CD (Luo et al., 2026), and CDGS (Mishra et al., 2026). For panoramic image generation, we compare with Multi-Diffusion (Bar-Tal et al., 2023) (the image-domain analogue of GSC), SyncDiffusion (Lee et al., 2023), and CDGS. For long video generation, we compare with Gen-L-Video (Wang et al., 2023) (the video-domain analogue of GSC), and CDGS. We also report CogVideoX-2B at its native short horizon of 50 frames as a video reference.



Figure 5: **Panoramic image generation.** The coarse-to-fine procedure maintains a consistent global style across distant chunks. Baselines show abrupt changes or gradual style drift over the panorama.

Compositional long-horizon robotic planning. Table 1 reports success rates on OGBench (Park et al., 2025). The medium and large `PointMaze` settings are saturated by existing compositional methods, so the harder Giant and Scene-Play tasks are more informative. On `PointMaze-Giant`, we achieve 96% success, improving over CDGS by 14%, suggesting that the scaffold construction stage helps resolve long-range inconsistency in maze-like trajectory composition. On `AntMaze-Giant`, we obtain 85%, comparable to CDGS at 84%. Here, the strongest compositional baselines already find reasonable global routes, and the remaining failures likely come from inverse-dynamics tracking (Luo et al., 2026). On `Scene-Play`, we improve over CDGS from 51% to 63%, where the gain suggests that the coarse scaffold stage provides a more stable scaffold for highly multimodal play data. Non-compositional baselines degrade sharply in the harder long-horizon settings. For example, on `PointMaze-Giant`, while compositional methods remain effective compared to non-compositional baselines obtain 0% success. Figure 4 shows a generated plan from CoFi on `PointMaze-Giant`. The scaffold construction stage captures the global route but may still produce wall-crossing artifacts. The subsequent refinement stages refine the plan into a globally aligned trajectory.

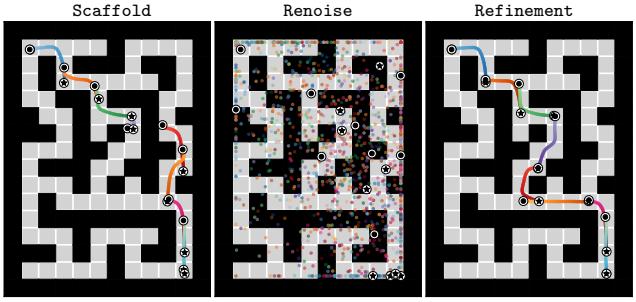

Figure 4: **Robotic Planning.** A coarse scaffold route is refined into a plausible task-level trajectory.

Panoramic image generation. Table 2 and Figure 5 summarize the panorama generation results. We achieve the best scores across all three metrics among the compared methods. In particular, relative to CDGS, our method reduces Intra-LPIPS from 0.65 to 0.48 and Intra-Style loss from 1.71 to 0.72, while also improving CLIP score from 32.67 to 33.18. These gains indicate that the generated patches remain more consistent in color, texture, and overall style across the full panorama, without sacrificing text-image alignment. Qualitatively, our method produces panoramas with a more stable global style and fewer abrupt transitions across patch boundaries. Unlike SyncDiffusion, which relies on an additional perceptual optimization loop during sampling, our method achieves strong coherence through the same coarse-to-fine procedure used in the planning experiments. This indicates that the proposed method improves global coherence while preserving prompt-relevant local details and semantic fidelity.

Table 2: **Quantitative results on panoramic image generation.** We compose 9 overlapping 512×512 local patches into a 512×4608 panorama and evaluate patch consistency and text alignment on generated panorama. Best results are highlighted.

	Style		Semantic
Method	LPIPS ↓	Style-L ↓	CLIP-S ↑
GSC (Multi-Diff)	0.72±0.08	2.96±0.24	31.77±2.14
Sync-Diff	0.60±0.05	1.36±1.07	32.40±1.85
CDGS	0.65±0.05	1.71±1.03	32.67±1.92
CoFi (Ours)	0.48±0.08	0.72±0.54	33.18±1.84

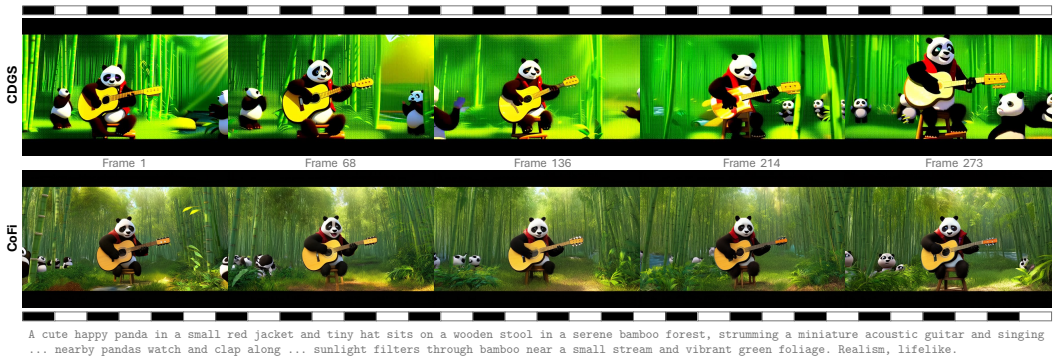

Figure 6: **Long video generation.** The coarse-to-fine procedure preserves subject appearance and scene structure over long temporal ranges. Baseline shows temporal drift along distant frames.

Table 3: **Quantitative results on long video generation.** We compose 50-frame short-horizon CogVideoX-2B videos into a 273-frame long-horizon video and evaluate Vbench on generated long-horizon videos. The 50-frame video serves as a reference. Best results are highlighted.

	Temporal		Frame	Semantic
Method	Subject ↑	Flickering ↑	Aesthetic ↑	Prompt ↑
CogVideoX 50frames	95.91	97.36	63.10	25.51
CogVideoX 273frames	90.24	98.44	49.44	21.78
GSC (Gen-L-Video)	89.51	96.89	60.12	25.13
CDGS	91.67	97.16	58.90	26.13
CoFi (Ours)	94.11	97.27	62.36	26.86

Long video generation. Table 3 and Figure 6 report the long video generation results. We improve subject consistency from 91.67 with CDGS to 94.11, approaching the native 50-frame CogVideoX reference at 95.91. This is the key metric in this setting, since temporal composition often fails through subject drift over long horizons. We also improve aesthetic quality from 58.90 with CDGS to 62.36, close to the native short-horizon reference at 63.10, and obtain the best prompt alignment among compositional baselines at 26.86. These results suggest that the refinement stage recovers local visual detail after the coarse global structure is fixed, rather than relying on expensive search throughout sampling. Temporal flickering is similar across compositional methods. Although the native 273-frame CogVideoX extension obtains the highest flickering score, this metric does not fully capture long-range subject drift or semantic inconsistency. As shown in Figure 6, the coarse-to-fine procedure better preserves subject appearance and scene layout across the full video.

Inference cost. Figure 7 compares the number of function evaluations (NFE) across the three domains. GSC is the cheapest because it uses local guidance without resampling or search, but the global structure remains under-specified. CDGS adds inner iterations, population search, and pruning to improve global coherence, which makes the cost grow quickly as the number of local plans increases. This is especially costly for panorama and video generation, where many high-dimensional chunks must be coordinated. In contrast, CoFi separates global and local refinement. The scaffold construction stage uses the same denoiser evaluations as GSC, and the refinement stage adds only $t^\star$ extra steps. The total cost is therefore $T + t^\star$, giving a small overhead over GSC while requiring 2–8× fewer function evaluations than CDGS.

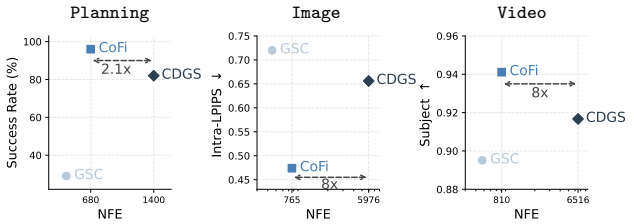

Figure 7: **NFE–performance Pareto comparison.** We compare performance against the NFE across the three domains. CoFi improves CDGS with **2–8× fewer NFE**, substantially reducing computational overhead while improving global coherence and local sample quality.

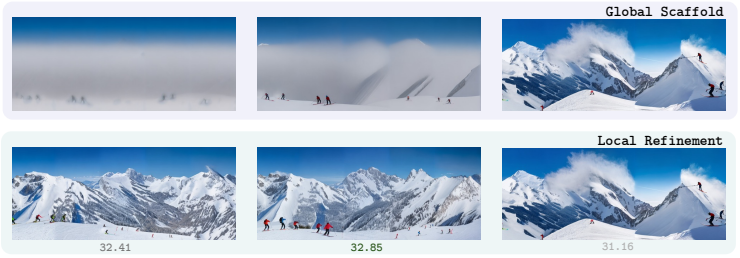

Figure 8: **Effect of scaffold schedule.** We compare strong scaffold alignment (left) with constant $\lambda_t = 0.8$, (middle) the scheduled λ_t in Figure 9, and (right) no scaffold alignment with constant $\lambda_t = 1$. The top row shows the coarse scaffold Y_0^{c}, and the bottom row shows the refinement Y_0^{r}. Numbers below the images report the CLIP score.

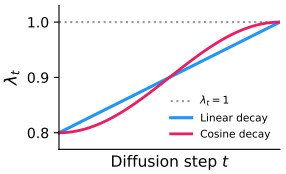

Figure 9: Smaller λ_t gives stronger scaffold alignment. We use stronger alignment at early diffusion steps to align local plans around a shared coarse mode, and gradually increase $\lambda_t \to 1$.

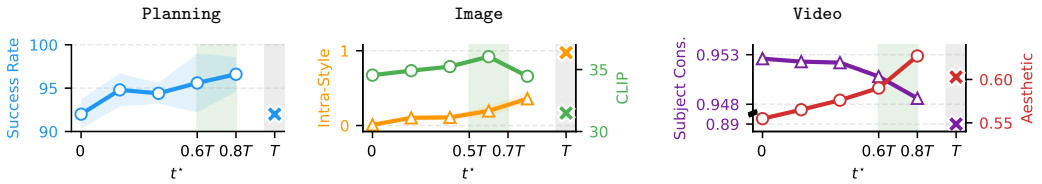

Figure 10: **Effect of refinement depth $t^\star$.** We vary $t^\star$ to measure the trade-off between *global structure preservation* and *local detail recovery*. Image metrics compare Intra-Style (*global*) with CLIP (*local*), while video metrics compare subject consistency (*global*) with frame-level aesthetic quality (*local*). The colored $\times$ denote generation from w/o scaffold alignment (*i.e.*, $Y^{\mathrm{r}}_{t^\star=T} \sim \mathcal{N}(0, I_d)$ in (16)) and intermediate $t^\star \in [0.5T, 0.8T]$ values form a `sweet spot` across domains.

Ablation Study. We study the role of the scaffold alignment coefficient λ_t in (14). Smaller λ_t pulls local denoised estimates more strongly toward the shared scaffold, whereas $\lambda_t = 1$ leaves them unchanged. As shown in Figure 8, a fixed $\lambda_t = 0.8$ over-smooths the scaffold, while $\lambda_t = 1$ removes scaffold construction and yields the lowest CLIP score of 31.16. We therefore use the scheduled λ_t shown in Figure 9, which applies stronger scaffold alignment at early denoising steps and gradually relaxes it. This schedule balances global coherence and local detail, producing a coherent coarse scaffold Y_0^{c} whose refined output Y_0^{r} achieves the highest CLIP score of 32.85.

We next examine the refinement depth $t^\star$ in (16), which controls the strength of the scaffold constraint during the second stage. Smaller $t^\star$ preserves the global structure Y_0^{c} but limits the the refinement stage to correct local artifacts. Larger $t^\star$ allows the local prior to modify local details more substantially, but increases the risk that the sample deviates from the global scaffold Y_0^{c}. As shown in Figure 10, both extremes are suboptimal. At $t^\star = 0$ (*i.e.*, w/o refinement in Algorithm 2), coarse-stage artifacts persist, showing degraded local quality (lower CLIP and aesthetic scores) despite strong global coherence, whereas at $t^\star = T$ (*i.e.*, w/o scaffold alignment in Algorithm 1), the global scaffold is discarded entirely, showing poor global coherence (higher style loss and lower subject consistency). Intermediate values $t^\star \in [0.5T, 0.8T]$ consistently yield the best balance between global structure preservation and local detail recovery across all three domains.

5 Conclusion

We introduced CoFi, an inference-time sampler that decomposes compositional generation into a consensus-based coarse stage and a structure-preserving refinement stage. By separating global structure formation from local detail recovery, we improve both global coherence and local quality over existing compositional baselines across robotic planning, panoramic image generation, and long video generation, while requiring 2–8$\times$ fewer denoiser evaluations.

Limitations. The quality of the composed output is bounded by the expressiveness of the pre-trained local prior, and the consensus schedule λ_t and refinement depth $t^\star$ require per-domain selection. Moreover, the current consensus applies uniform weighting to all local plans, which may be suboptimal when segments differ in structural importance.

Acknowledgments

This work was partly supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) with a grant funded by the Ministry of Science and ICT (MSIT) of the Republic of Korea in connection with the Global AI Frontier Lab International Collaborative Research Agreement (No. RS-2024-00469482 & RS-2024-00509279) and Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No.RS-2019-II190075, Artificial Intelligence Graduate School Program(KAIST)). U. M, J. C, and Y. C are supported by NSF Award 2409016.

References

Michael Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *Journal of Machine Learning Research*, 26(209):1–80, 2025.

Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. *arXiv preprint arXiv:2302.08113*, 2023.

Yilun Du and Leslie Pack Kaelbling. Position: Compositional generative modeling: A single model is not all you need. In *Forty-first International Conference on Machine Learning*, 2024.

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

Dibya Ghosh, Abhishek Gupta, Ashwin Reddy, Justin Fu, Coline Manon Devin, Benjamin Eysenbach, and Sergey Levine. Learning to reach goals via iterated supervised learning. In *International Conference on Learning Representations*, 2021. https://openreview.net/forum?id=rALA0Xo6yNJ.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024.

Subin Kim, Seoung Wug Oh, Jui-Hsien Wang, Joon-Young Lee, and Jinwoo Shin. Tuning-free multi-event long video generation via synchronized coupled sampling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6418–6429, 2025.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit q-learning. *arXiv preprint arXiv:2110.06169*, 2021.

Yuseung Lee, Kunho Kim, Hyunjin Kim, and Minhyuk Sung. Syncdiffusion: Coherent montage via synchronized joint diffusions. *Advances in Neural Information Processing Systems*, 36:50648–50660, 2023.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. https://openreview.net/forum?id=PqvMRDCJT9t.

Yunhao Luo, Utkarsh Aashu Mishra, Yilun Du, and Danfei Xu. Generative trajectory stitching through diffusion composition. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. https://openreview.net/forum?id=VCTt5DXiBe.

Corey Lynch, Mohi Khansari, Ted Xiao, Vikash Kumar, Jonathan Tompson, Sergey Levine, and Pierre Sermanet. Learning latent plans from play. In *Conference on robot learning*, pages 1113–1132. Pmlr, 2020.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022. https://openreview.net/forum?id=aBsCjcPu_tE.

Utkarsh Aashu Mishra, Shangjie Xue, Yongxin Chen, and Danfei Xu. Generative skill chaining: Long-horizon skill planning with diffusion models. In *Conference on Robot Learning*, pages 2905–2925. PMLR, 2023.

Utkarsh Aashu Mishra, David He, Yongxin Chen, and Danfei Xu. Compositional diffusion with guided search for long-horizon planning. In *The Fourteenth International Conference on Learning Representations*, 2026. https://openreview.net/forum?id=b8avf4F2hn.

Seohong Park, Dibya Ghosh, Benjamin Eysenbach, and Sergey Levine. Hiql: Offline goal-conditioned rl with latent states as actions. *Advances in Neural Information Processing Systems*, 36:34866–34891, 2023.

Seohong Park, Kevin Frans, Benjamin Eysenbach, and Sergey Levine. OGBench: Benchmarking offline goal-conditioned RL. In *The Thirteenth International Conference on Learning Representations*, 2025. https://openreview.net/forum?id=M992mjgKzI.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

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

Fu-Yun Wang, Wenshuo Chen, Guanglu Song, Han-Jia Ye, Yu Liu, and Hongsheng Li. Gen-l-video: Multi-text to long video generation via temporal co-denoising. *arXiv preprint arXiv:2305.18264*, 2023.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan.Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. In *The Thirteenth International Conference on Learning Representations*, 2025. https://openreview.net/forum?id=LQzN6TRFg9.

Jonathan S Yedidia, William Freeman, and Yair Weiss. Generalized belief propagation. *Advances in neural information processing systems*, 13, 2000.

Jonathan S Yedidia, William T Freeman, and Yair Weiss. Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Transactions on information theory*, 51(7): 2282–2312, 2005.

Qinsheng Zhang, Jiaming Song, Xun Huang, Yongxin Chen, and Ming-Yu Liu. Diffcollage: Parallel generation of large content with diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10188–10198. IEEE, 2023.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

Yixin Zhang, Yunhao Luo, Utkarsh Aashu Mishra, Woo Chul Shin, Yongxin Chen, and Danfei Xu. Compositional visual planning via inference-time diffusion scaling. In *The Fourteenth International Conference on Learning Representations*, 2026. https://openreview.net/forum?id=EEONns7ae4.

A Derivation of the Closed-Form Coarse Estimate

We derive the closed-form solution of the coarse alignment objective in (12). For notational clarity, we drop the time index t and write $\hat{X}^i := \hat{X}^i_{0|t}$ for the Tweedie estimate of the i-th local plan. Let $\gamma > 0$ denote the regularization strength and define

$$f(Z^1, \dots, Z^N) \;=\; \sum_{i=1}^{N} \Big[\|Z^i - \hat{X}^i\|^2 \;+\; \gamma \, \|Z^i - \bar{Z}\|^2 \Big], \qquad \bar{Z} := \tfrac{1}{N} \textstyle\sum_{i=1}^{N} Z^i. \tag{17}$$

Proof of (13). Since (17) is a sum of convex quadratics in $(Z^1, \dots, Z^N)$, the objective is strictly convex and admits a unique global minimizer. We derive the first-order optimality conditions below.

Let δ_{ij} denote the Kronecker delta. Differentiating (17) with respect to Z^j and applying the chain rule with $\partial \bar{Z} / \partial Z^j = \frac{1}{N} I_m$ yields

$$\frac{1}{2} \frac{\partial f}{\partial Z^j} \;=\; \big(Z^j - \hat{X}^j\big) \;+\; \gamma \sum_{i=1}^{N} \big(Z^i - \bar{Z}\big) \Big(\delta_{ij} - \tfrac{1}{N}\Big). \tag{18}$$

By construction, $\sum_{i=1}^{N} (Z^i - \bar{Z}) = \mathbf{0}$. Expanding the summation in (18) and applying this identity, we obtain

$$\frac{1}{2} \frac{\partial f}{\partial Z^j} \;=\; \big(Z^j - \hat{X}^j\big) \;+\; \gamma \big(Z^j - \bar{Z}\big). \tag{19}$$

Setting (19) to zero and solving for Z^j:

$$(1+\gamma)\, Z^j \;=\; \hat{X}^j + \gamma\, \bar{Z} \qquad \Longrightarrow \qquad Z^j \;=\; \lambda\, \hat{X}^j \;+\; (1-\lambda)\, \bar{Z}, \quad \text{where } \lambda := \tfrac{1}{1+\gamma}. \tag{20}$$

Summing both sides of (20) over $j = 1, \dots, N$ and dividing by N gives $\bar{Z} = \lambda\, \bar{X} + (1-\lambda)\, \bar{Z}$, where $\bar{X} := \frac{1}{N} \sum_{i=1}^{N} \hat{X}^i$. Rearranging yields $\lambda\, \bar{Z} = \lambda\, \bar{X}$, and hence

$$\bar{Z}^\star \;=\; \bar{X}. \tag{21}$$

Substituting (21) into (20) produces the minimizer $Z^{j,\star} = \lambda\, \hat{X}^j + (1-\lambda)\, \bar{X}$, which coincides with (13) upon restoring the time index t and setting $\lambda_t := 1/(1+\gamma_t)$. □

B Experimental Details

We adopt the experimental protocol of Mishra et al. (2026) and use the same pretrained diffusion models, datasets, and evaluation metrics. Below we provide the hyperparameter settings specific to our method.

B.1 Compositional Long-Horizon Robotic Planning

Task description. We evaluate on OGBench (Park et al., 2025), which provides offline datasets for goal-conditioned robotic tasks. The agent must navigate from a given start state to a goal state, and the offline dataset contains only short trajectory segments that do not individually solve the evaluation tasks. Successful planning therefore requires composing multiple short segments into a coherent long-horizon plan that reaches the goal while respecting environment constraints.

- `PointMaze` (`Stitch`): A 2D point agent navigates through a maze from a start position to a goal position. The observation and goal spaces are both 2D (XY position). We evaluate on `Medium`, `Large`, and `Giant` map sizes, where increasing scale requires more composed segments and longer navigation paths through the maze.
- `AntMaze` (`Stitch`): An 8-DoF ant robot navigates through the same maze layouts as `PointMaze`. Although the full observation is 29-dimensional, planning is performed in the 2D XY subspace, and a learned inverse-dynamics model maps the planned waypoints back to the full action space for execution.

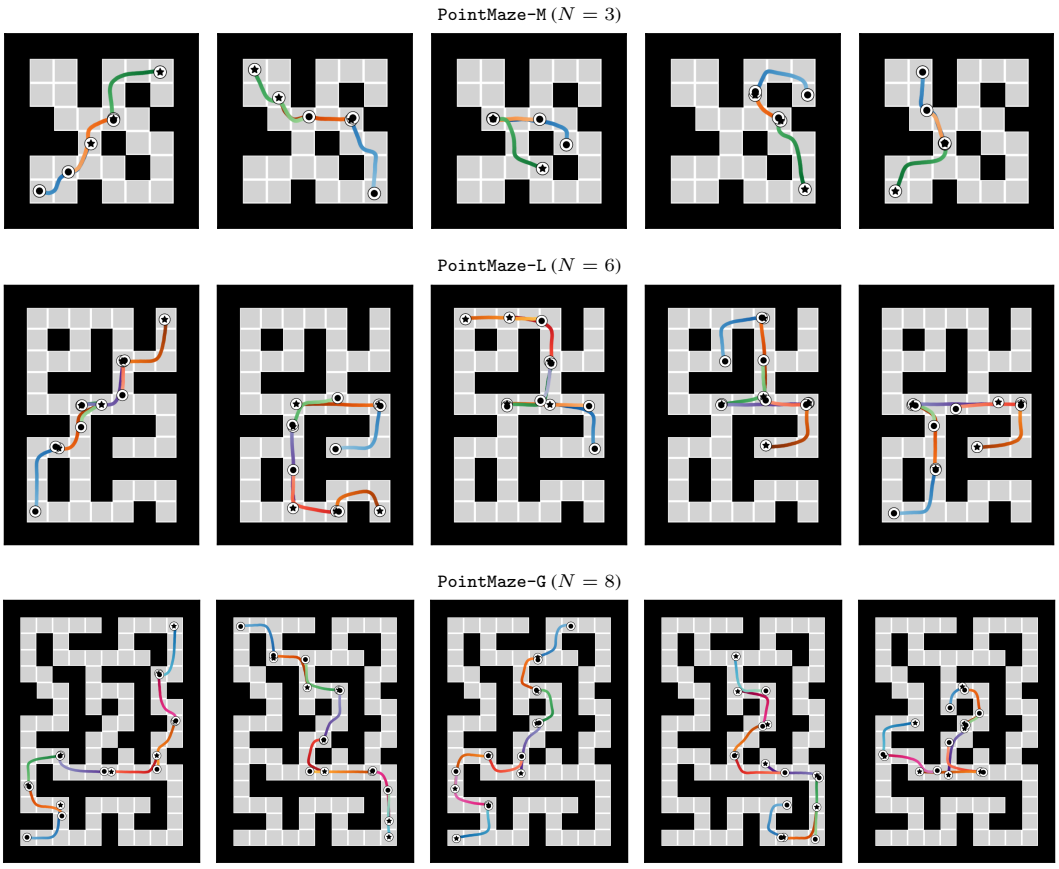

Figure 11: **PointMaze trajectory examples.** Each row shows five evaluation tasks at increasing maze scale. Each color denotes a distinct local plan segment.

- `Scene` (`Play`): A tabletop manipulation task with a robot gripper and multiple objects. The planning space is 14-dimensional, selected from the full 40-dimensional observation: gripper 3D position (indices 12–14), two object 3D positions (indices 19–21, 26–29), object orientations (indices 32–33, 36, 38), and contact states (indices 8–11). The environment defines 5 predefined tasks, each specifying a distinct start and goal configuration that requires the gripper to reach, grasp, and reposition one or more objects on the table. Unlike the maze tasks where the agent must find a collision-free path through a static environment, `Scene-Play` requires coordinating gripper motion with object interaction, and the offline play dataset contains unstructured exploration rather than goal-directed demonstrations. We evaluate 20 trials per task for a total of 100 episodes per seed. During execution, contact dimensions (indices 8–11) are snapped to binary values $\{0, 1\}$ to match the discrete nature of contact events in the training data.

Table 4 summarizes the per-environment configuration, including the observation and goal dimensions, local plan horizon and overlap length, default number of composed segments N, replanning distance threshold, and maximum environment steps.

Guidance cost. The local guidance cost in (11) is the overlap mean-squared error between the predicted clean boundaries of adjacent segments:

$$\hat{r}_t(X_t^i, X_t^{i+1}) = -c \left\| \mathtt{R}(\hat{X}_{0|t}^i) - \mathtt{L}(\hat{X}_{0|t}^{i+1}) \right\|^2, \tag{22}$$

where $c = 5.0$. This gradient-based guidance is used in both the coarse stage (Algorithm 1) and the refinement stage (Algorithm 2). After each denoising step, the composed global trajectory is obtained by exponentially blending adjacent segments in their overlap regions (Luo et al., 2026). Concretely,

Table 4: Environment specifications for robotic planning.

Environment	obs dim	goal dim	horizon	overlap	N	replan thres	max steps
`PointMaze-Medium`	2	2	160	56	3	1.0	1000
`PointMaze-Large`	2	2	160	56	6	1.0	1000
`PointMaze-Giant`	2	2	160	56	8	1.0	1000
`AntMaze-Medium`	29	2	160	56	3	4.0	1000
`AntMaze-Large`	29	2	160	56	6	4.0	1000
`AntMaze-Giant`	29	2	160	56	9	4.0	2000
`Scene-Play`	40	14	128	48	8	2.0	1500

for an overlap of length L, the blended trajectory at position $\ell \in \{0, \dots, L-1\}$ is

$$\texttt{Blend}(\ell) \;=\; w(\ell) \cdot \texttt{R}(X^i) \;+\; \big(1 - w(\ell)\big) \cdot \texttt{L}(X^{i+1}), \qquad w(\ell) \;=\; \frac{e^{-\beta \ell/(L-1)} - e^{-\beta}}{1 - e^{-\beta}}, \tag{23}$$

where $\beta = 2.0$ is used.

Diffusion model and hyperparameters. We use the same pretrained 1D U-Net diffusion models from Luo et al. (2026); Mishra et al. (2026), trained on the OGBench offline datasets with 512 diffusion steps. Each local plan covers a trajectory segment of 160 (maze) or 128 (scene) timesteps, and N overlapping segments are composed to cover the full start-to-goal trajectory. At inference time, we use $T = 100$ DDIM steps with stochasticity $\eta = 1.0$. For the coarse stage, we use a ramped λ_t schedule: λ_t starts at $\lambda_0 = 0.2$ and linearly increases to 1.0. The refinement depth is set to $t^\star = 0.7T$ for `Maze`. We generate a batch of 40 candidate trajectories per planning problem and select the one with the lowest overlap MSE.

Evaluation protocol. We report success rates over 100 evaluation episodes averaged across 5 random seeds. Each evaluation episode samples a random start-goal pair from the OGBench evaluation set. An episode is considered successful if the agent reaches within a threshold distance of the goal within the maximum number of environment steps (see Table 4). For non-compositional baselines (GCBC, GCIVL, GCIQL, HIQL), we use the results reported in Mishra et al. (2026).

Replanning and execution. For maze tasks, we follow Luo et al. (2026) and use distance-based replanning: the agent replans when it deviates beyond the threshold distance (see Table 4) from the current planned trajectory. Actions are extracted from the planned waypoints using a learned inverse-dynamics model that maps consecutive observation pairs to low-level actions. For `Scene-Play`, the agent replans up to 10 times with a distance threshold of 2.0, and the planned trajectory is executed for up to 1500 environment steps per episode.

Per-task trajectory examples. Figures 11 and 12 show composed trajectories generated by our method on each maze environment across different evaluation tasks, and Figure 13 shows rollout keyframes for the five Scene-Play manipulation tasks. Each color denotes a distinct local plan segment, with overlapping boundaries between adjacent segments. The start position is marked with $\circ$ and the goal with $\star$. As the maze scale increases from `Medium` to `Giant`, the number of composed segments N grows accordingly, and the coarse stage becomes increasingly important for maintaining a globally coherent route.

B.2 Bethe Approximation for Compositional Diffusion

Factor graph formulation. Recall from Section 2 that the global plan $Y = [x^1, x^2, \dots, x^{2N+1}]$ is composed of N overlapping local plans $X^i = [x^{2i-1}, x^{2i}, x^{2i+1}]$, where each adjacent pair (X^i, X^{i+1}) shares a boundary factor x^{2i+1}. This structure naturally defines a factor graph: each local plan X^i corresponds to a *factor node* with local prior $p_t(X_t^i)$, and each shared boundary x_t^{2i+1} corresponds to a *variable node* connecting the two adjacent factors. The joint distribution over the

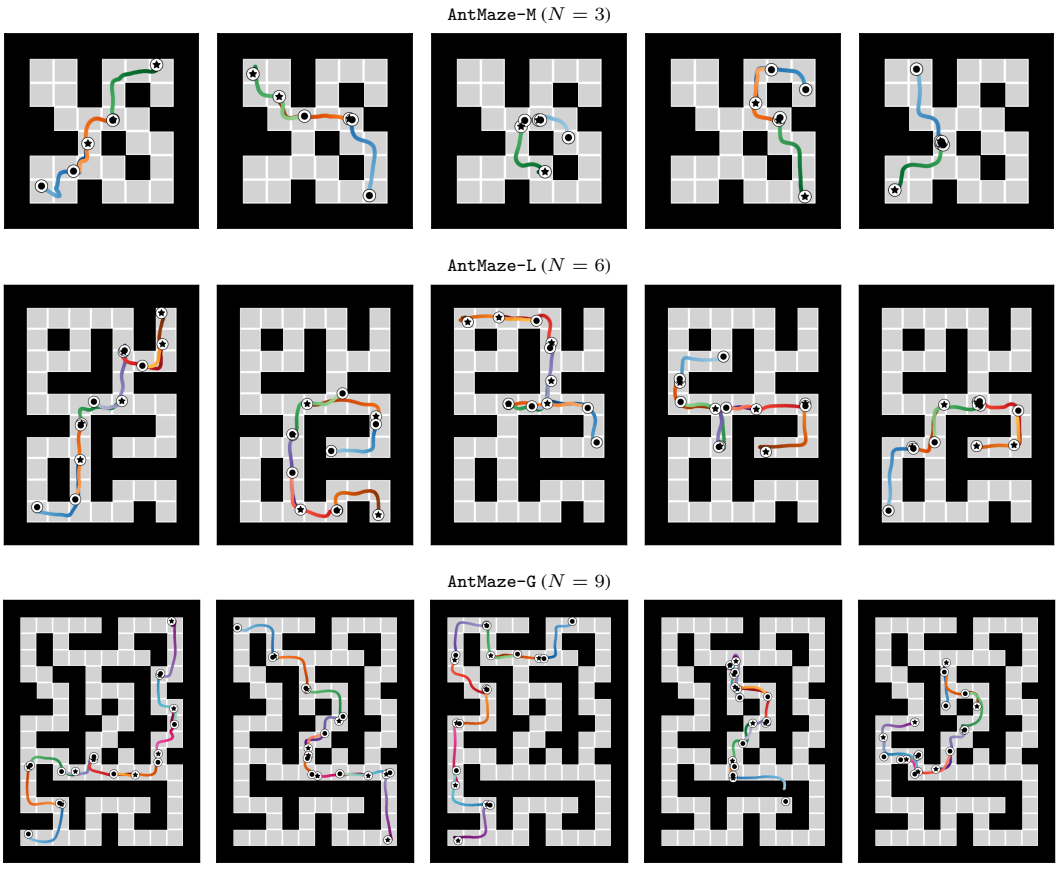

Figure 12: **AntMaze trajectory examples.** Each row shows five evaluation tasks at increasing maze scale. Each color denotes a distinct local plan segment.

global plan factorizes over the factor graph as

$$p_t(Y_t) \;\propto\; \prod_{i=1}^{N} p_t(X_t^i). \tag{24}$$

Since the local plans overlap, this product double-counts the shared boundary variables. The Bethe free energy (Yedidia et al., 2000, 2005) corrects for this overcounting by subtracting the marginal log-densities at the shared boundaries:

$$\log p_t(Y_t) \;\approx\; \sum_{i=1}^{N} \log p_t(X_t^i) \;-\; \sum_{i=1}^{N-1} \log p_t(x_t^{2i+1}). \tag{25}$$

The first sum aggregates the local log-densities (one per factor node), and the second sum removes the overcounted overlap marginals (one per shared variable node). This is the standard Bethe approximation applied to the chain-structured factor graph.

Differentiating (25) with respect to the shared boundary x_t^{2i+1} yields the compositional score:

$$\nabla_{x_t^{2i+1}} \log p_t(Y_t) \;\approx\; \underbrace{\nabla \log p_t\big(\mathtt{R}(X_t^i) \mid X_t^i\big)}_{\texttt{from left factor}} + \underbrace{\nabla \log p_t\big(\mathtt{L}(X_t^{i+1}) \mid X_t^{i+1}\big)}_{\texttt{from right factor}} - \nabla \log p_t(x_t^{2i+1}), \tag{26}$$

where $\mathtt{R}(X_t^i) = \mathtt{L}(X_t^{i+1}) = x_t^{2i+1}$ is the shared boundary. When the overlap marginal $p_t(x_t^{2i+1})$ is approximated as uniform or absorbed into the local scores, this reduces to the form in (10): an average of the two conditional scores from the left and right factors.



Figure 13: **Scene-Play rollout examples.** Each row shows keyframes from a successful rollout for one of the five predefined manipulation tasks. The robot gripper must reach, grasp, and reposition objects on the table according to the task-specific goal configuration.

B.3 Panoramic Image Generation

Task description. Given a text prompt, the task is to generate a wide panoramic image (512×4608) that maintains consistent style, color, and semantic content across the full spatial extent. The pretrained Stable Diffusion 2.0 (Rombach et al., 2022) model generates 512×512 images natively, so producing a panorama requires composing $N = 9$ overlapping patches with 50% spatial overlap.

Guidance. Local consistency across adjacent patches is enforced via the Bethe approximation in (10). Concretely, in the shared spatial overlap between patches X^i and X^{i+1}, the fused denoised estimate is obtained by ramp blending rather than uniform averaging. At spatial position ℓ within an overlap of width W:

$$\hat{X}_{0|t}^{\texttt{fused}}(\ell) \;=\; w(\ell) \cdot \hat{X}_{0|t}^{i} \;+\; \big(1 - w(\ell)\big) \cdot \hat{X}_{0|t}^{i+1}, \qquad w(\ell) \;=\; 1 - \frac{\ell}{W}, \tag{27}$$

which linearly transitions from the left patch to the right. This ramp blending produces smoother boundary transitions than equal-weight averaging. Global coherence is achieved through the coarse mean regularization in (13).

Hyperparameters. We use $T = 50$ DDIM steps with classifier-free guidance scale 7.5. For the coarse stage, we set $\lambda_0 = 0.2$ with a linear ramp schedule that increases λ_t to 1.0. The clean estimate $\hat{X}_{0|t}^{i}$ is obtained from the fused (overlap-averaged) denoised output. We use sampled directional noise in the coarse stage, meaning each local plan receives independently sampled noise for the DDIM direction term rather than sharing a common noise realization. For the refinement stage, we

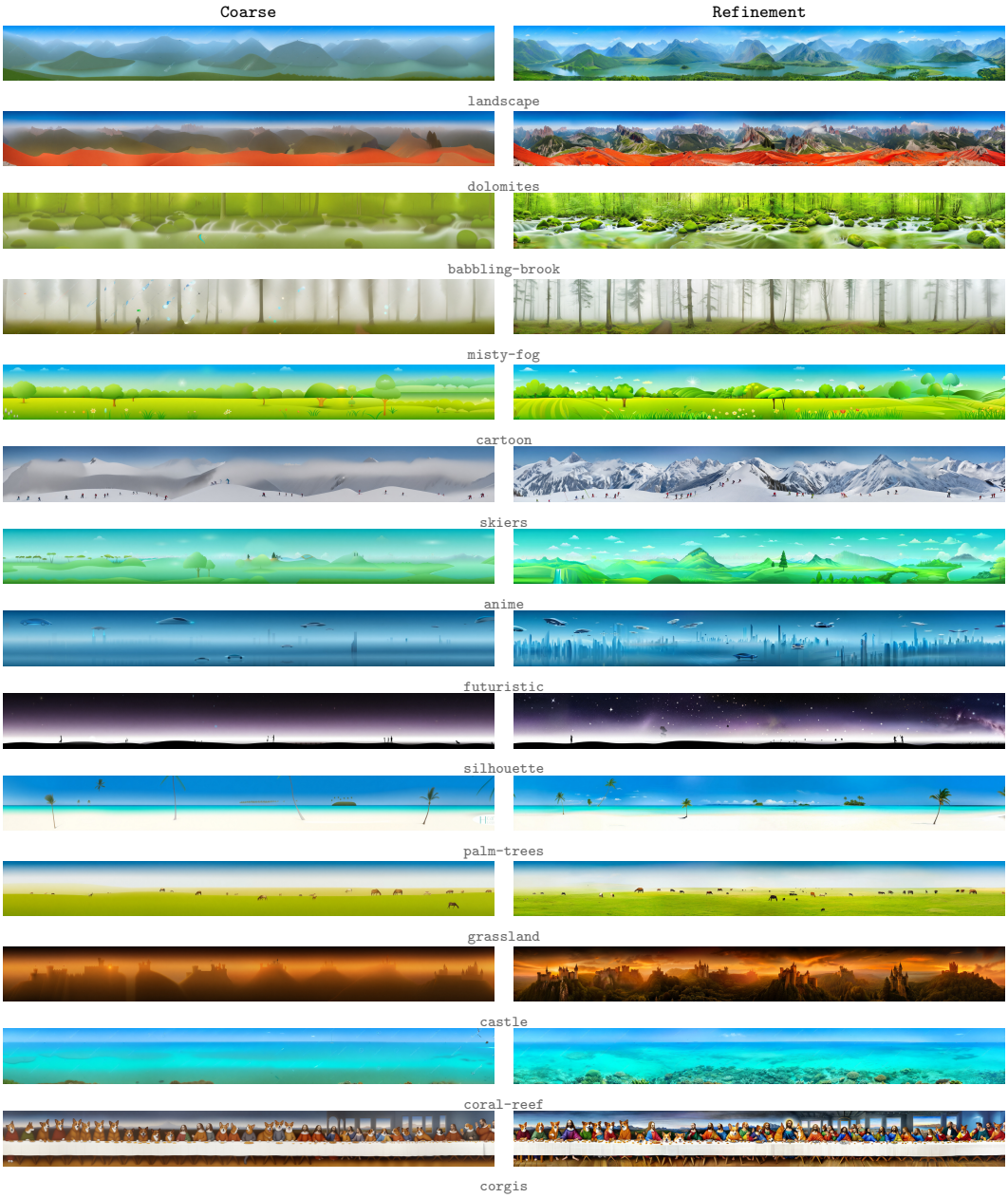

Figure 14: **Panoramic image results.** Per-prompt coarse scaffold and refined output (full prompts in Appendix B.8).

set $t^\star = 0.7T$ (refinement begins from step 15) with stochasticity $\eta = 1.0$ and ramp blending. We generate 5 random seeds per prompt and report the mean and standard deviation across seeds.

Metrics. *Intra-LPIPS* (Zhang et al., 2018) measures the mean pairwise LPIPS perceptual distance between all pairs of non-overlapping 512×512 crops within a single panorama, capturing local style consistency across distant spatial regions. *Intra-Style loss* (Gatys et al., 2016) computes the mean pairwise Gram-matrix distance between VGG features (layers `relu1_2`, `relu2_2`, `relu3_3`, `relu4_3`) of these crops, measuring global style coherence. *CLIP score* (Radford et al., 2021) evaluates text-image alignment between the prompt and the full generated panorama using the ViT-B/32 CLIP model.

Per-prompt results. We evaluate on 14 text prompts with 5 random seeds each. Figure 14 shows the coarse scaffold and the refined output for all prompts.

B.4 Long Video Generation

Task description. Given a text prompt, the task is to generate a temporally coherent long video (273 frames at 720p) that preserves subject identity and scene structure over a much longer horizon than the pretrained model natively supports. CogVideoX-2B (Yang et al., 2025) generates 50-frame clips, so producing a long video requires composing $N = 9$ temporal chunks with 50% overlap.

Guidance. As in the image domain, local temporal consistency across adjacent chunks is enforced via the Bethe approximation in (10) with ramp blending in the shared temporal overlap. At temporal position ℓ within an overlap of L frames:

$$\hat{X}_{0|t}^{\texttt{fused}}(\ell) \;=\; w(\ell)\cdot \hat{X}_{0|t}^{i} \;+\; \big(1-w(\ell)\big)\cdot \hat{X}_{0|t}^{i+1}, \qquad w(\ell) \;=\; 1-\frac{\ell}{L}. \tag{28}$$

Global temporal coherence is achieved through the coarse mean regularization in (13).

Hyperparameters. We use $T = 50$ DDIM steps with classifier-free guidance scale 6.0. The coarse stage uses $\lambda_0 = 0.2$ with the same linear ramp schedule as the image setting. Unlike the image setting, we use noise-bar directional noise (rather than independently sampled noise) in the coarse stage: all local chunks share a common noise realization in the DDIM direction term, which we find critical for maintaining temporal coherence across chunk boundaries. The coarse mean decay follows a cosine schedule$\lambda_t \;=\; \lambda_0 \cdot \cos\!\big(\frac{\pi\, t}{2T}\big)$, which provides a smoother transition from strong to weak regularization compared to linear decay. For the refinement stage, we set $t^\star = 0.7T$ with stochasticity $\eta = 1.0$.

Metrics. We use VBench (Huang et al., 2024) to evaluate long video quality across four dimensions. *Subject consistency* measures the mean pairwise DINO feature similarity of the main subject across uniformly sampled frame pairs, capturing whether the subject's appearance is preserved over the full video duration. *Temporal flickering* evaluates frame-to-frame visual stability by measuring the mean absolute pixel difference between consecutive frames. *Aesthetic quality* applies a learned aesthetic predictor to individual frames and reports the mean score. *Prompt alignment* measures text-video correspondence via CLIP similarity between the text prompt and uniformly sampled frames.

Per-prompt results. We evaluate on 8 text prompts with 8 random seeds each. Figure 15 shows sampled frames from the coarse scaffold and the refined output for all prompts.

B.5 Baseline Configurations

For all compositional baselines, we use the implementations and hyperparameters from CDGS[1][2] (Mishra et al., 2026). GSC (Mishra et al., 2023) (and its image/video analogues MultiDiffusion (Bar-Tal et al., 2023) and Gen-L-Video (Wang et al., 2023)) applies local guidance without resampling or search. SyncDiffusion[3] (Lee et al., 2023) adds a perceptual optimization loop at each denoising step. CDGS (Mishra et al., 2026) uses population-based resampling with likelihood-based pruning across inner iterations.

B.6 Compute Resources

All experiments are conducted on a single NVIDIA A6000 GPU (48GB VRAM). For robotic planning, composing $N = 9$ segments with $T = 100$ DDIM steps requires approximately 2 GB of GPU memory. For panoramic image generation, composing a 512×4608 panorama from 9 patches with $T = 50$ steps requires approximately 16 GB. For long video generation, composing 273 frames from 9 temporal chunks with $T = 50$ steps requires approximately 23.5 GB

[1] https://github.com/UtkarshMishra04/CDGS_imgvideo under Apache 2.0 license.
[2] https://github.com/UtkarshMishra04/CDGS_ogbench under MIT license.
[3] https://github.com/KAIST-Visual-AI-Group/SyncDiffusion under MIT license.

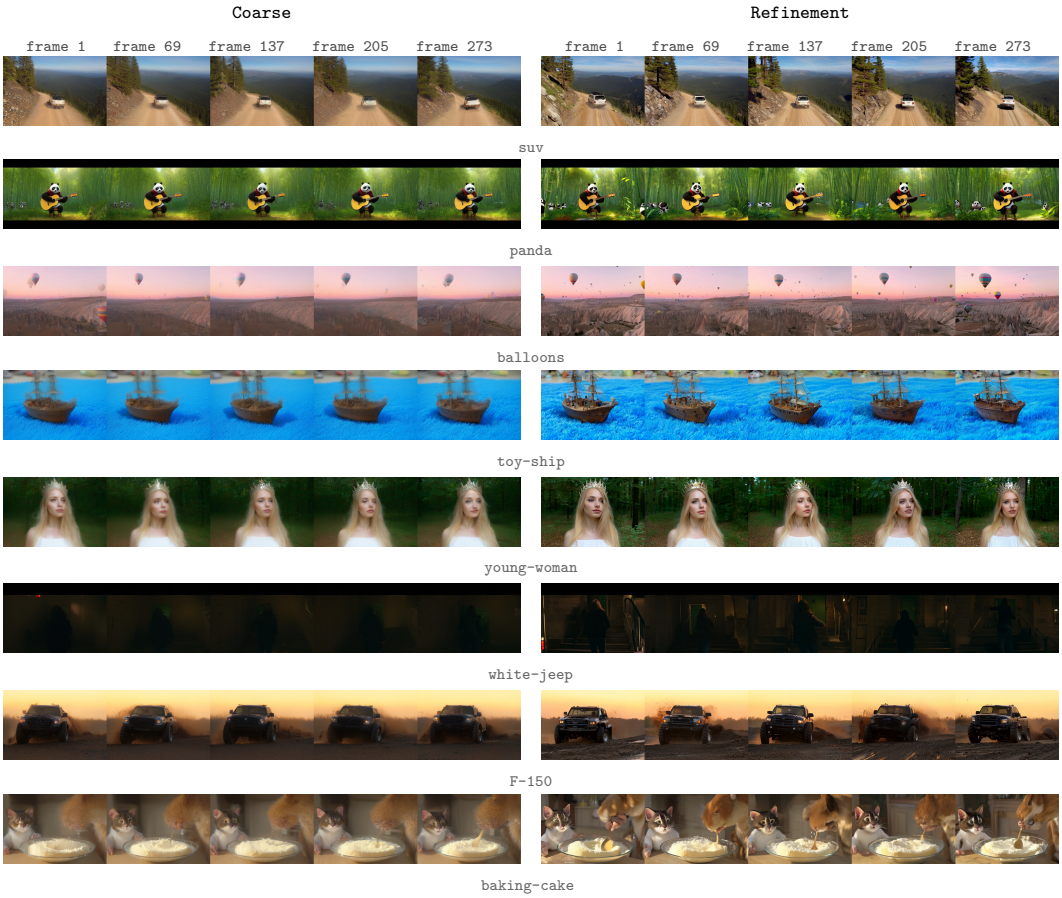

Figure 15: **Long video results.** Per-prompt coarse scaffold and refined output. Each strip shows frames 1, 69, 137, 205, 273 of the 273-frame composed video (full prompts in Appendix B.8).

B.7 NFE Analysis

We compare the number of function evaluations (NFE) across methods. Let T denote the total diffusion steps and N the number of local plans.

- **GSC** : Each denoising step requires one forward pass per local plan. Total NFE $= T \cdot N$.
- **CDGS**: Each denoising step involves K inner iterations with a population of size P. Total NFE $= K \cdot T \cdot P \cdot N$, where typical values are $K \in [5, 10]$ and $P \in [4, 8]$.
- **Ours**: The coarse stage runs T steps and the refinement stage runs $t^\star$ additional steps, each with one forward pass per local plan. Total NFE $= (T + t^\star) \cdot N$. With $t^\star = 0.3T$, this gives $1.3T \cdot N$.

The ratio of CDGS to our NFE is $\frac{K \cdot P}{1.3}$. For typical CDGS settings ($K = 3$, $P = 4$), this gives approximately $9.2\times$, consistent with the 2–8$\times$ reduction reported in the main text (the exact ratio depends on domain-specific CDGS configurations).

B.8 Full Prompt Lists

Panoramic image prompts. We evaluate on the following 14 text prompts:

- `landscape`: "A beautiful landscape with mountains and a river"
- `dolomites`: "A photo of the dolomites with red lava flowing through the valley"
- `babbling-brook`: "A photo of lush forest with a babbling brook"
- `misty-fog`: "A photo of a forest with a misty fog"
- `cartoon`: "Cartoon panorama of spring summer beautiful nature"

- `skiers`: “A photo of a snowy mountain peak with skiers”
- `anime`: “Natural landscape in anime style illustration”
- `futuristic`: “Skyline of a futuristic city with flying cars”
- `silhouette`: “Silhouette wallpaper of a dreamy scene with shooting stars”
- `palm-trees`: “A beach with palm trees”
- `grassland`: “A photo of a grassland with animals”
- `castle`: “A cinematic view of a castle in the sunset”
- `coral-reef`: “A photo of a beautiful ocean with coral reef”
- `corgis`: “last supper with cute corgis”

Long video prompts. We evaluate on the following 8 text prompts:

- `suv`: “The camera follows behind a white vintage SUV with a black roof rack as it speeds up a steep dirt road surrounded by pine trees on a steep mountain slope, dust kicks up from its tires, the sunlight shines on the SUV as it speeds along the dirt road, casting a warm glow over the scene. The dirt road curves gently into the distance, with no other cars or vehicles in sight. The trees on either side of the road are redwoods, with patches of greenery scattered throughout. The car is seen from the rear following the curve with ease, making it seem as if it is on a rugged drive through the rugged terrain. The dirt road itself is surrounded by steep hills and mountains, with a clear blue sky above with wispy clouds.”
- `panda`: “A cute happy panda, dressed in a small, red jacket and a tiny hat, sits on a wooden stool in a serene bamboo forest. The panda’s fluffy paws strum a miniature acoustic guitar, producing soft, melodic tunes, move hands, singings. Nearby, a few other pandas gather, watching curiously and some clapping in rhythm. Sunlight filters through the tall bamboo, casting a gentle glow on the scene. The panda’s face is expressive, showing concentration and joy as it plays. The background includes a small, flowing stream and vibrant green foliage, enhancing the peaceful and magical atmosphere of this unique musical performance.”
- `balloons`: “A group of colorful hot air balloons take off at dawn in Cappadocia, Turkey. Dozens of balloons in various bright colors and patterns slowly rise into the pink and orange sky. Below them, the unique landscape of Cappadocia unfolds, with its distinctive ‘fairy chimneys’ – tall, cone-shaped rock formations scattered across the valley. The rising sun casts long shadows across the terrain, highlighting the otherworldly topography.”
- `toy-ship`: “A detailed wooden toy ship with intricately carved masts and sails is seen gliding smoothly over a plush, blue carpet that mimics the waves of the sea. The ship’s hull is painted a rich brown, with tiny windows. The carpet, soft and textured, provides a perfect backdrop, resembling an oceanic expanse. Surrounding the ship are various other toys and children’s items, hinting at a playful environment. The scene captures the innocence and imagination of childhood, with the toy ship’s journey symbolizing endless adventures in a whimsical, indoor setting.”
- `young-woman`: “A young woman with beautiful and clear eyes and blonde hair standing and white dress in a forest wearing a crown. She seems to be lost in thought, and the camera focuses on her face. The video is of high quality, and the view is very clear. High quality, masterpiece, best quality, highres, ultra-detailed, fantastic.”
- `white-jeep`: “A woman walks away from a white Jeep parked on a city street at night, then ascends a staircase and knocks on a door. The woman, wearing a dark jacket and jeans, walks away from the Jeep parked on the left side of the street, her back to the camera; she walks at a steady pace, her arms swinging slightly by her sides; the street is dimly lit, with streetlights casting pools of light on the wet pavement; the camera follows the woman from behind as she walks up a set of stairs towards a building with a green door; she reaches the top of the stairs and turns left, continuing to walk towards the building; she reaches the door and knocks on it with her right hand.”
- `F-150`: “At sunset, a modified Ford F-150 Raptor roared past on the off-road track. The raised suspension allowed the huge explosion-proof tires to flip freely on the mud, and the mud splashed on the roll cage.”
- `baking-cake`: “A cat and a dog baking a cake together in a kitchen. The cat is carefully measuring flour, while the dog is stirring the batter with a wooden spoon. The kitchen is cozy, with sunlight streaming through the window.”

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims in the abstract and introduction are well aligned with the scope of the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The conclusion section discusses the limitations of the presented work.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: The paper does not present theoretical results, so this question is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides sufficient details to reproduce the main experimental results, including the experimental setup, evaluation metrics, and implementation details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The supplemental material provides anonymized code and instructions to reproduce the main experimental results, including environment setup, data access, and execution commands.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: The proposed method is training-free, so training details such as optimizer choice or training schedules are not applicable. The paper specifies the relevant inference-time and evaluation details, including the datasets, evaluation protocols, hyperparameters, and baseline settings needed to understand the results.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We report results over five random seeds and include error bars defined as the standard deviation across seeds where applicable.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper reports that the experiments were run on a single NVIDIA RTX A6000 GPU, which provides the main compute resource required for reproduction.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The research conforms to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The paper focuses on a training-free inference-time algorithmic method and does not include a separate discussion of positive and negative societal impacts.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release new datasets or trained models with high misuse risk. It proposes a training-free inference-time method using existing pretrained models.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper properly credits the pretrained models, and baseline implementations used in the experiments.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not introduce new datasets or trained models as assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing experiments or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve human subjects or study participants.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not used as an important, original, or non-standard component of the core methodology. Any use of LLMs, if any, was limited to writing, editing, and did not affect the scientific content of the work.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.