CoPlace: Joint Subject Placement and Empty Scene Planning for Sequential Image Generation

Anonymous authors
Paper under double-blind review

Abstract

Multi-modal Procedure Planning (MPP) generates textual and visual steps towards accomplishing a specific goal. While recent developments in large language models have empowered effective textual reasoning and planning, generating coherent visual procedures remains challenging due to the requirements of subject consistency, state transition coherence, and subject-background harmony. Existing methods typically treat these problems separately, leading to inevitable trade-offs.

To address these challenges, we propose Transformation, Plan then Place (TPP), a novel multi-stage framework for sequential image generation. TPP decomposes the generation process into three stages: subject state transformation, joint subject layout and empty-scene planning, and subject placement. We also introduce CoPlace, a joint generation model, where layout planning and scene generation iteratively constrain each other during the denoising process. Furthermore, we propose a space-gating mechanism that suppresses text leakage within subject placement regions, encouraging the generation of a semantically empty subject placeholder that is compatible with global scene context.

Extensive experiments on HD-EPIC, LayoutSAM, and HiCo-7K demonstrate that our methods achieve superior performance in improving layout planning, image quality, and geometric compatibility between subjects and backgrounds. Our approach provides a flexible solution for generating coherent multi-step visual procedures while effectively balancing subject fidelity, text alignment and scene controllability.



Figure 1: A example of Multi-modal procedure planning.



Figure 2: Fidelity VS Text Alignment.

1 Introduction

Drastic development in artificial intelligence (Achiam et al., 2023; Bai et al., 2022) has paved the way for powerful applications such as AI assistants(Mialon et al., 2024) and agents (Agashe 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.

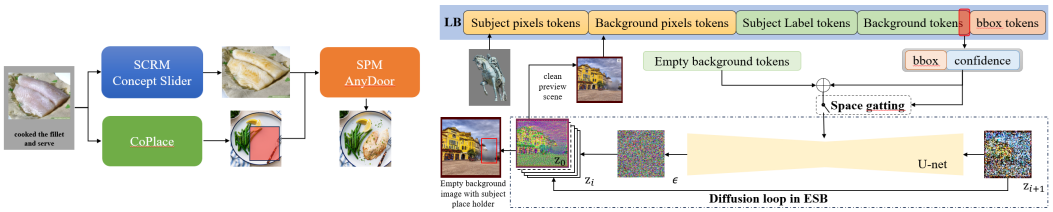
2025; Xie et al., 2024), enabling knowledge acquisition in minutes and thereby shifting the bottleneck to effectively applying knowledge to solving real-world problems. Increasing attention has been paid to Multi-modal Procedure Planning (MPP) (Wang et al., 2026; Ilaslan et al., 2025) focusing on generating a sequence of actions in multiple modalities to achieve a specific goal. The versatile reasoning capabilities of LLMs enable plausible textual procedural generation through dedicated prompt engineering. However, intricate requirements in visual branch are often underexplored, even though they are crucial to deliver satisfactory user experience. The generated image sequence is expected to satisfy: (1) **subject consistency** across different steps, (2) **coherent state transitions** between consecutive steps, and (3) **subject–background harmony** in terms of spatial layout, scale, occlusion, and overall visual compatibility. As shown in Fig. 1, the turkey should remain identical throughout the entire image sequence with suitable state transitions, from raw, to covered, then cooked. What is more, the turkey should also naturally interact with surrounding objects, such as orange, oil or pan, in terms of spatial arrangement and physical contact.

Existing works in customized image generation (Zhang et al., 2025b; Cao et al., 2025), semantic-controllable image generation (Gandikota et al., 2024; Park et al., 2023; Haas et al., 2024; Wang et al., 2025; Baumann et al., 2025) and layout-aware image generation (Srivastava et al., 2025; Liu et al., 2026; He et al., 2025) have made significant progress towards satisfying different subsets of these aspects. However, a comprehensive solution that simultaneously satisfies all these requirements remains elusive and challenging. Firstly, the rade-off between image fidelity and text alignment limits single model performance, as fine-grained reconstruction of reference subjects often sacrifices controllability (Zhang et al., 2025b). As shown in Fig. 2, based on reference text, the cinnamon sticks must be placed within surrounding wooden top, white napkin and pumpkin decoration while maintaining identical features as the reference image. Secondly, a suitable subject placement depends on the affordances of the surrounding objects and good placement needs reasoning about relative scale, free space, support relations, occlusion, containment, etc. Unlike existing methods which independently fabricate subject layout before scene construction, we argue that subject placement should be jointly planned with scene generation, since the proper subject position inherently depends on the surrounding environment.

Aiming to address these challenges, we propose Transformation, Plan then Place (TPP), a novel multi-stage framework and CoPlace, a joint subject layout and empty-scene generation model. Considering that the reference subjects across different steps often exhibit varying states, poses and interaction, disentangling subject's state transition, empty scene generation with subject layout planning and subsequent subject placement enables focused optimization towards a single specific goal at each stage and thus alleviates the trade-off problem and facilitates plug-and-play flexibility. Specifically, the framework comprises three modules: (1) a State-Conditioned Rendering Module (SCRM), (2) the CoPlace, and (3) a Subject Placement Module (SPM). SCRM applies selected Concept Sliders (Gandikota et al., 2024) to perform controllable subject-state transformation. As shown in Fig. 3b, CoPlace is designed to reduce subject occlusion or semantically incompatibility in placeholder regions, allowing scene composition and subject placement to bidirectionally constrain each other before subject rendering. We also introduce a novel space-gating mechanism to suppress background text leakage into bbox regions. By replacing text guidance within bbox regions with empty conditions, the in-bbox regions are encouraged to be constructed through global scene context rather than explicit text clues. As a result, the reserved regions can remain semantically subject-free while naturally blending into surrounding background. Finally, utilizing plug and play object swapping modules, such as AnyDoor (Chen et al., 2024), SPM seamlessly integrates refined subjects into empty backgrounds. We also conduct extensive experiments on HD-EPIC, LayoutSAM and HiCo-7K datasets. Experimental results demonstrate our superior performance on metrics in terms of layout planning, image quality and subject-scene compatibility.

Our contributions are fourfold:

1. We propose a novel multi-stage framework for sequential image generation;
2. We formulate the problem of joint subject layout and empty background generation;
3. We introduce a joint subject layout and empty scene generation model to enhance subject-scene harmony;
4. We conduct extensive experiments to validate the effectiveness of our proposed framework and model, demonstrating superior performance in terms of subject consistency, state transition coherence, and subject-scene harmony.



(a) TPP consists of three modules: SCRM, CoPlace, and SPM. SCRM performs state-conditioned transitions; CoPlace jointly plans the empty scene and subject placeholder; finally, SPM blends the refined subject into the empty scene.

(b) Overview of CoPlace. The Layout Branch (LB) takes the clean preview scene generated by the Empty Scene Branch (ESB) and predicts a bounding box, while ESB uses the bounding box to generate a clean preview scene. The spatial gating mechanism suppresses explicit textual-clue leakage within the bounding-box region.

Figure 3: Overview of the proposed TPP framework and its CoPlace module.

2 Related Work

Customized Image Generation. Customized image generation aims to synthesize images with specific subjects adapted to given text conditions. This line of work can be approximately divided into two categories: test-time fine-tuning methods and tuning-free methods. Textual Inversion (Gal et al., 2022) encrypt subject identities into textual tokens. DreamBooth (Ruiz et al., 2023) fine-tunes the diffusion model to associate the reference subject to a unique identifier. These methods are computationally expensive at inference time. Later approaches focus on reducing computational cost. ELITE (Wei et al., 2023), BLIP-Diffusion Li et al. (2023a) and IP-Adapter (Ye et al., 2023) encode reference subject into textual feature or visual feature spaces that is injected into diffusion model through attention without subject-specific model tuning. While customized methods exhibit excellent performance on preserving subject fidelity, constructing desired images with exact textual semantic are relatively difficult due to the inherent trade-off between subject fidelity and text alignment.

Semantic-controllable Image Generation. This line of methods steer images towards selected semantics by manipulating different spaces in diffusion models, including h-space (Kwon et al., 2022; Haas et al., 2024), pixel space (Park et al., 2023) or noisy latent space (Wang et al., 2025), text embedding space (Baumann et al., 2025), parameter space (Gandikota et al., 2024), middle activation space (Cywiński & Deja, 2025) etc. Concept Slider (Gandikota et al., 2024) or SAeUron (Cywiński & Deja, 2025) can effectively construct images with desired attributes, nevertheless subject identities are not necessarily preserved.

Layout-Aware Image Generation. Early approaches (Li et al., 2023b; Zheng et al., 2023; Feng et al., 2023) followed a two-stage paradigm: the layout was inferred, followed by final image generation. Recent models, PlanGEN (He et al., 2025) and ATLAS (Liu et al., 2026), unify layout planning and image generation in one model. However, their causal autoregressive mechanism still conducts layout planning before image rendering. Different from existing methods, we propose a dual model with reciprocal adaptation to layout planning and empty scene generation enabling flexible subject rendering in the downstream plug and play module.

Multi-modal Procedural Planning (MPP). MPP studies how to produce a coherent sequence of steps to solve a high-level goal in multi-modal formats, such as texts or images. TIP (Lu et al., 2024) improved the interaction between text steps and image steps. VG-TVP (Ilaslan et al., 2025) extended to text-video planning. MMPlanner (Tabassum et al., 2025) considers object state reasoning before planning. However, relatively few works have focused on improving visual planning quality.

3 Method

3.1 Problem Formulation

We consider the task of joint subject layout and empty background generation (JSLEBG). Given a reference subject image I_{ref} and a text description of a scene T, we aim to generate a bbox B

that represents a placement holder for subsequent subject rendering and an empty background image S that leaves the space slot for subject insertion surrounded by necessary background objects. Formally, the generation problem can be formulated as:

$$(B, S_\phi) \sim P_{\text{CoPlace}}(B, S_\phi \mid T, I_{\text{ref}}) \tag{1}$$

JSLEBG constructs necessary layout and background before subject rendering, thus enables separating itself with the upstream task state conditioned subject adaptation and the downstream task subject placement.

3.2 Overview

Given a reference subject image I_{ref} and procedural text steps $T_{\text{ref}}^{(1)}, \ldots, T_{\text{ref}}^{(N)}$, our goal is to generate a sequence of images $I^{(1)}, \ldots, I^{(N)}$ that faithfully reproduce the reference subject in suitable states across steps while maintaining plausible interactions with the textually aligned surrounding scene.

As shown in Fig. 3a, in order to alleviate the trade-off between subject fidelity and text alignment, we propose **Transformation, Plan then Place (TPP)** that is composed of three modules: (1) a State-Conditioned Rendering Module (SCRM), which transforms the reference subject into proper state at each procedural step; (2) CoPlace, the joint subject placement and empty background generation module, and (3) a Subject Placement Module (SPM), which inserts the transformed subject into the planned region of the generated empty scene. TPP follows transform, plan then place paradigm. Formally, for procedural step i, TPP decomposes image generation as

$$\tilde{I}_{\text{st}}^{(i)} = \mathcal{T}\left(I_{\text{ref}}, T^i\right) \tag{2}$$

$$\left(B^i, S_\emptyset^i\right) \sim P_{\text{CoPlace}}\left(B^i, S_\emptyset^i \mid T^i, \tilde{I}_{\text{st}}^{(i)}\right) \tag{3}$$

$$I^{(i)} = \mathcal{P}\left(S_\emptyset^i, B^i, \tilde{I}_{\text{st}}^{(i)}, \right) \tag{4}$$

where $\mathcal{T}$, P_{CoPlace}, $\mathcal{P}$ denote SCRM, CoPlace and SPM respectively. To simplify the notation, the superscript i consistently denotes the i-th time step. $\tilde{I}_{\text{st}}^i$ and I_{ref} represents the reference subject image in suitable state and original subject image respectively. T^i is the text description. B^i and $S_\emptyset$ denotes subject bounding box and empty scene. I^i denote the final constructed image.

In this work, we implement the State-conditioned Rendering Module and the Subject Placement Module with fine-tuned PlanGEN He et al. (2025) and AnyDoor Chen et al. (2024) respectively.

3.3 State-Conditioned Rendering Module

In the MPP setting, an identical subject in different steps often exhibits substantial changes in diverse states. For example, a food ingredient may transition from raw to cooked. What is worse, the subject could perform complex interactions with numerous background objects, potentially causing occlusion and incompatibility. SCRM focuses on state adaption while maintaining the subject identity, leaving scene planning and interaction to the SPM. To be specific, we apply Concept Slider (Gandikota et al., 2024), which provides a lightweight and efficient framework for continuous attribute control, enabling smooth adjustment of semantic attributes while disentangling the target concept from other irrelevant visual factors. The Concept Slider transfers reference subject I_{ref} into suitable state $\tilde{I}_{\text{st}}$,

3.4 CoPlace: Joint Empty-Scene and Subject-Layout Generation

Unlike existing methods that conduct layout planning and whole image generation sequentially, we propose CoPlace that performs reciprocal optimization on both subject layout planning and empty background generation. As shown in Fig. 3b, CoPlace is primarily comprised of the Layout Branch (**LB**) and the Empty Scene Branch (**ESB**). The former one predicts subject bbox with confidence score based on the reference image, **preview background image**, a subject label and empty scene description. The later performs DDIM sampling and outputs an empty scene image conditioned on empty background description as well as additional **subject bbox** and **confidence score** from LB. Aiming to suppress subject leakage and prevent occlusion, we also design a novel Space-Gating Mechanism (**SGM**).

Scene-Aware Placement Generation. We implement LB by adapting PlanGen(He et al., 2025), an autoregressive vision language model, into a layout predictor given empty background description, preview scene image and visual features from ESB, formulated as:

$$L_\phi(B, c|\tilde{I}_{\text{st}}, L_{sub}, S_{pre}, S_{text}) \tag{5}$$

where B denotes the target bounding box, c denotes the confidence score of B, $\tilde{I}_{st}$ and L_{sub} are the subject image in proper state and the subject label respectively, S_{pre} and S_{text} represent the estimated empty scene image and textual description respectively. These features are flattened, inputted into the backbone sequentially and then encoded into a global feature token that is leveraged to predict a bbox and the confidence by the newly added bbox head and quality head. The bbox head is a pre-norm two-layer MLP with GELU activation, predicting 4-dimensional bounding box coordinates. The quality head is simply a LayerNorm and one-layer MLP that output a confidence score. Notably, bbox tokens generated by next token prediction in the VLM backbone serve as auxiliary supervision signal rather than explicit coordinates.

Layout-Conditioned Scene Generation. ESB is implemented as a standard Stable Diffusion model 1.5 except that it is additionally conditioned on B and c from LB through a Spatial Text Gate Mechanism. Given an empty scene description $T^t_\emptyset$, estimated bounding box B^t and confidence score c^t, it gradually generates an empty scene image through DDIM sampling:

$$z_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\hat{z}_0 + \sqrt{1-\bar{\alpha}_{t-1}}\epsilon_\theta(z_t, t, (T^t_\emptyset, B^t, c^t)). \tag{6}$$

Space-Gating Mechanism. To suppress text leakage into the preservation region, we introduce a *Space Gate* that removes text and space guidance from the region inside the bbox. At each denoising step, we perform two UNet forward passes: (1) a conditional prediction $\epsilon_{\text{full}} = \text{UNet}(z_t, t, (T_\emptyset, B^i, c^i))$, conditioned on the empty background text, the current layout bounding-box tokens and corresponding confidence score, (2) an unconditional prediction $\epsilon_{\text{uncond}} = \text{UNet}(z_t, t, \emptyset)$. Given the predicted bbox, we construct a soft spatial mask $M \in [0,1]^{H\times W}$ in the latent space, where M approaches 1 inside the target region and 0 outside. We then spatially modulate classifier-free guidance as

$$\epsilon_{\text{gated}} = \epsilon_{\text{uncond}} + \lambda(1-M) \odot (\epsilon_{\text{full}} - \epsilon_{\text{uncond}}), \tag{7}$$

where λ denotes the classifier-free guidance scale and $\odot$ represents element-wise multiplication. Consequently, regions outside the bounding box receive standard conditioned guidance, while the conditional residual is progressively suppressed inside the target box, reducing the tendency to synthesize salient content in the region reserved for subject placement, which is thus forced to be constructed by attending to surrounding visual features rather than explicit text clues. A soft mask is used around the box boundary to ensure a smooth spatial transition and avoid visible artifacts caused by hard gating.

Reciprocal Evolution. Unlike existing approaches that perform layout planning independently before image rendering, we formulate subject layout planning and empty scene rendering as mutually conditioned generation process. Heuristically, we activate communication between LB and ESB after the first 20% of DDIM sampling steps. Thereafter, the bbox and estimated preview empty scene are iteratively updated and exchanged every 10% of the denoising process, allowing the two branches to progressively refine each other.

Subject Placement Module (SPM). As shown in Eq. equation 4, after CoPlace produces the empty scene $S_\emptyset$ and subject layout B, SPM composites the refined subject $\tilde{I}_{\text{st}}$ into the layout region. In this paper, we implement this plug-and-play SPM using AnyDoor (Chen et al., 2024).

3.5 Training Objectives

We optimize CoPlace in three stages applying $\mathcal{L}_{\text{LB}}$ to LB and $\mathcal{L}_{\text{SEB}}$ to ESB. LB is first trained to predict a geometrically accurate and well-calibrated placement of the reference subject. Then we freeze LB and train ESB using ground truth bbox in the warmup stage and estimated bbox from LB in the main stage sequentially.

3.5.1 LAYOUT PREDICTION OBJECTIVE.

We consider leveraging coordinate regression ($\mathcal{L}_{\text{l1}}$), overlap ($\mathcal{L}_{\text{iou}}$), scale ($\mathcal{L}_{\text{area}}$), textual grounding ($\mathcal{L}_{\text{grd}}$) and confidence calibration ($\mathcal{L}_{\text{qual}}$) for LB optimization. Formally,

$$\mathcal{L}_{\text{layout}} = \lambda_{\text{l1}}\mathcal{L}_{\text{l1}} + \lambda_{\text{iou}}\mathcal{L}_{\text{iou}} + \lambda_{\text{area}}\mathcal{L}_{\text{area}} + \lambda_{\text{grd}}\mathcal{L}_{\text{grd}} + \lambda_{\text{qual}}\mathcal{L}_{\text{qual}}. \tag{8}$$

where $\mathcal{L}_{\text{layout}}$ denotes the layout loss and the λ_* terms are weighting coefficients.

Bounding-box Regression and Text Grounding. $\mathcal{L}_{\text{l1}}$ supervising the normalized box coordinates is a smooth-$L1$ objective (Girshick, 2015). The textual grounding objective $\mathcal{L}_{\text{grd}}$ applies the standard autoregressive cross-entropy loss supervising ground truth bbox coordinates, as an auxiliary objective to preserve grounding capability of underlying VLM while additionally strengthening spatial prediction.

Geometric Supervision. We also encourage estimated bbox to maintain natural geometric features. Let $b^* = (x_1^*, y_1^*, x_2^*, y_2^*)$ denote the ground-truth bounding box and $\hat{b} = (\hat{(}x_1), \hat{(}y_1), \hat{(}x_2), \hat{(}y_2))$ the box predicted by the layout branch. IoU loss promotes exact alignment with target region as a whole:

$$\mathcal{L}_{\text{iou}} = 1 - \text{IoU}(\hat{b}, b^*), \tag{9}$$

where IoU measures the Intersection over union between $\hat{b}$ and b^*. The area consistency objective constrains the $\hat{b}$ to maintain an appropriate scale with respect to b^*:

$$A(b) = (x_2 - x_1)(y_2 - y_1) \tag{10}$$

$$\mathcal{L}_{\text{area}} = \left|\log\left(A(\hat{b}) + \epsilon\right) - \log\left(A(b^*) + \epsilon\right)\right|, \tag{11}$$

where ϵ is a small constant for numerical stability.

Confidence Calibration. We supervise the predicted quality score using binary cross-entropy (BCE) with the IoU as a soft target $q^* = \text{sg}[\text{IoU}(\hat{b}, b^*)]$. $\text{sg}[\cdot]$ denotes the stop gradient operation. Let $\hat{q}$ denote output of confidence head,

$$\mathcal{L}_{\text{qual}} = \text{BCE}(\hat{q}, q^*), \tag{12}$$

3.5.2 SPATIALLY WEIGHTED DIFFUSION OBJECTIVE.

On top of the Space Gating Mechanism mentioned above, we also assign extra weights on reserved region. Given that $M(p) \in [0, 1]$ denotes the ground-truth subject mask at latent location p, $a = A(b^*)$ denotes the normalized area of the ground-truth bbox, we define a spatial weighting function as

$$w(p) = 1 + \lambda_{\text{region}} M(p), \tag{13}$$

$$\lambda_{\text{region}} = \lambda_{\text{subject}} \frac{1 - a}{a}. \tag{14}$$

λ_{subject} is a hyperparameter and the area-adaptive factor $\frac{1-a}{a}$ scales the region weight inversely with the target area, assigning greater emphasis on smaller bbox regions. Then, based on the original Eq. equation 18, the spatial weighted diffusion loss is defined:

$$\mathcal{L}_{\text{SEB}} = \frac{\sum_p w(p) \left\|\epsilon_\theta(x_t, t, c)_p - \epsilon_p\right\|_2^2}{\sum_p w(p)}. \tag{15}$$

4 EXPERIMENTS

4.1 TRAINING DATASETS AND IMPLEMENTATION DETAILS

As shown in Tab 3, we utilize LayoutSAM (Zhang et al., 2025a), HiCo-7K (Cheng et al., 2024) and HD-EPIC (Perrett et al., 2025). LayoutSAM is collected from diverse real-world images covering various everyday scenes; HiCo-7K is composed of complex real-world scene images involving multiple objects and spatial relationships; HD-EPIC is collected from first-person kitchen scenarios capturing human activities and object interactions. We crop subjects from original images and employ Big-LaMa to generate empty background reference by inpainting. We also leverage GPT-5.6 to

Table 1: Full-image generation results on three datasets.

Dataset	Method	Image Quality				Subject Consistency	
		Pick↑	CLIP↑	FID↓	IS↑	DINO-S↑	CLIP-S↑
LayoutSAM	IP-Adapter	18.941	0.630	66.269	18.743	0.5822	0.7849
	Lay-Your-Scene	20.756	0.805	36.116	19.304	0.2960	0.7172
	PlanGen	21.258	0.810	34.108	**20.743**	0.3639	0.7200
	PlanGen+AnyDoor	19.664	0.730	55.105	14.308	**0.5875**	0.7858
	CoPlace+AnyDoor	**21.322**	**0.836**	**33.994**	19.385	0.5808	**0.7871**
HiCo-7K	IP-Adapter	18.092	0.644	156.626	13.952	**0.6084**	0.7853
	Lay-Your-Scene	18.232	0.693	130.599	14.397	0.3228	0.7258
	PlanGen	19.693	**0.738**	131.630	15.702	0.2973	0.6728
	PlanGen+AnyDoor	17.835	0.629	147.454	10.957	0.6026	0.7831
	CoPlace+AnyDoor	**19.775**	0.685	**130.462**	**15.827**	0.5977	**0.7875**
HD-EPIC	IP-Adapter	18.405	0.582	269.648	7.821	0.4850	0.7826
	Lay-Your-Scene	19.759	0.737	229.086	**10.518**	0.2234	0.6530
	PlanGen	19.546	0.736	131.629	7.714	0.1843	0.6062
	PlanGen+AnyDoor	18.321	0.607	**93.528**	3.762	**0.4859**	0.7924
	CoPlace+AnyDoor	**19.788**	**0.785**	122.786	8.227	0.4852	**0.7937**

Table 2: Empty-scene generation results on three datasets.

Dataset	Method	Image Quality				Empty-region Quality	
		Pick↑	CLIP↑	FID↓	IS↑	Overlay↓	Leakage↓
LayoutSAM	PlanGen_tuned	19.582	0.694	**33.108**	12.971	0.049	0.199
	CoPlace	**20.541**	**0.740**	50.277	**15.104**	**0.046**	**0.192**
	CoPlace-SG	20.040	0.739	58.957	14.653	0.072	0.224
	CoPlace-SG-SE	20.034	0.738	58.531	14.184	0.079	0.223
HiCo-7K	PlanGen_tuned	17.645	0.570	**106.933**	10.520	0.238	0.107
	CoPlace	**19.880**	0.669	136.593	**13.662**	**0.199**	**0.054**
	CoPlace-SG	19.803	**0.671**	138.934	13.093	0.244	0.089
	CoPlace-SG-SE	19.756	0.631	138.887	13.155	0.094	0.252
HD-EPIC	PlanGen_tuned	17.970	0.629	**91.197**	3.735	**0.006**	**0.043**
	CoPlace	**18.788**	**0.643**	124.595	6.356	0.007	0.081
	CoPlace-SG	18.763	0.641	129.464	**6.415**	0.009	0.091
	CoPlace-SG-SE	18.747	0.639	129.161	6.345	0.010	0.095

exclude subject-related information from global image description. Then, we reorganize the fields. Each final sample consists of a full image, a global image description, a subject image, a subject label, an empty background reference and an empty background description.

The LB and ESB branches adopt Janus-Pro-1B from PlanGen and Stable Diffusion 1.5, respectively. We first train the LB branch for 30,000 steps with a batch size of 2 and gradient accumulation steps of 4. Then, we warm up the ESB branch for 20,000 steps using ground-truth bounding boxes under the same batch size and gradient accumulation settings. Finally, we fine-tune the ESB branch using the bounding boxes predicted by the LB branch for an additional 1,200 steps, with a batch size of 2 and gradient accumulation steps of 2. All experiments are conducted on a single NVIDIA RTX 5090 GPU using the AdamW optimizer.

4.2 Subject-Consistent Image Generation

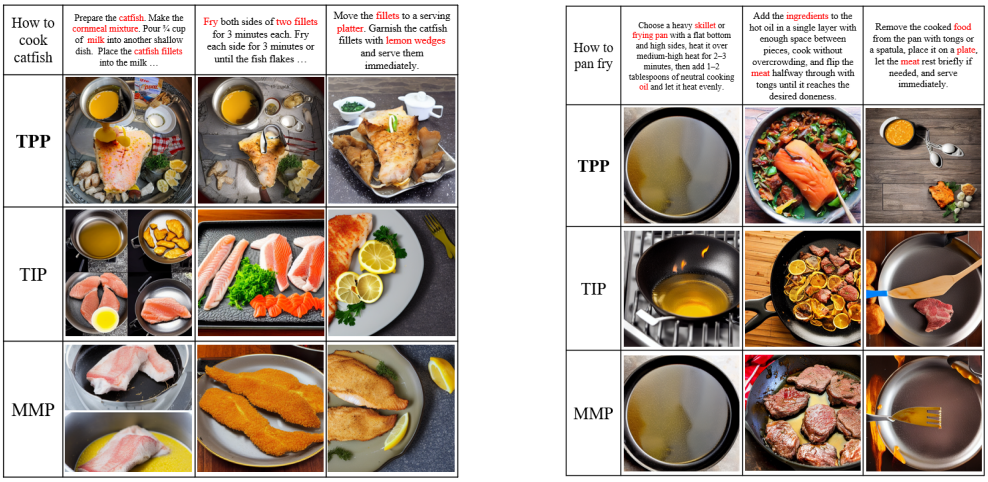
Settings. We compare our work CoPlace+AnyDoor with: 1) IP-Adapter Ye et al. (2023), 2) Lay-Your-Scene Srivastava et al. (2025), 3) PlanGen He et al. (2025) and 4) PlanGen+AnyDoor Chen et al. (2024) where PlanGen is fine-tuned to conduct joint empty scene and subject placeholder

Table 3: Training dataset statistics.

Dataset	Train	Dev/Val	Test	Total
LayoutSAM	23,203	2,842	2,817	28,862
HiCo-7K	7,178	896	937	9,011
HD-EPIC	10,338	925	920	12,183
Total	**40,719**	**4,663**	**4,674**	**50,056**

Table 4: Dataset-wise human preference rates (%). Best results are shown in bold.

Dataset	Method	Fidelity	Alignment	Harmony
LayoutSAM	**CoPlace+AnyDoor**	**74.1**	**46.3**	**53.1**
	PlanGen+AnyDoor	23.5	39.5	20.4
	Lay-Your-Scene	2.5	14.2	26.5
HiCo-7K	**CoPlace+AnyDoor**	**71.9**	**56.3**	**62.5**
	PlanGen+AnyDoor	26.0	22.9	11.5
	Lay-Your-Scene	2.1	20.8	26.0
HD-EPIC	**CoPlace+AnyDoor**	**69.0**	**50.0**	**66.7**
	PlanGen+AnyDoor	23.8	40.5	16.7
	Lay-Your-Scene	7.1	9.5	16.7



(a) Catfish-cooking example.

(b) Pan-frying example.

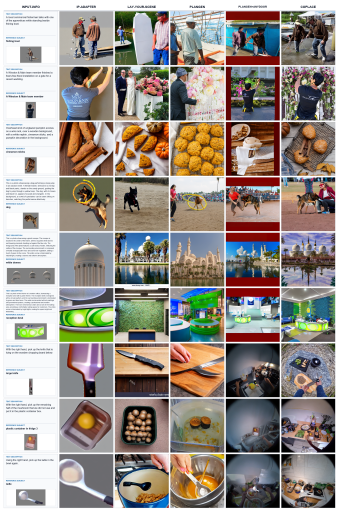
Figure 4: Qualitative comparisons of TPP with TIP and MPP on two procedural image-generation examples.

prediction. We leverage 1) PickScoreKirstain et al. (2023), 2) CLIPScore, 3) FID Heusel et al. (2017), 4) InceptionScore Salimans et al. (2016), 5) OverlayScore Li et al. (2026), 6) LeakageScore which is the cosine similarity between the grounding bbox and the predicted bbox by DINO Liu et al. (2024), 7) DINO-S, the cosine similarity between the DINO features extracted from the ground truth and predicted bboxes, and 8) CLIP-S, similar to DINO-S but using CLIP features instead.

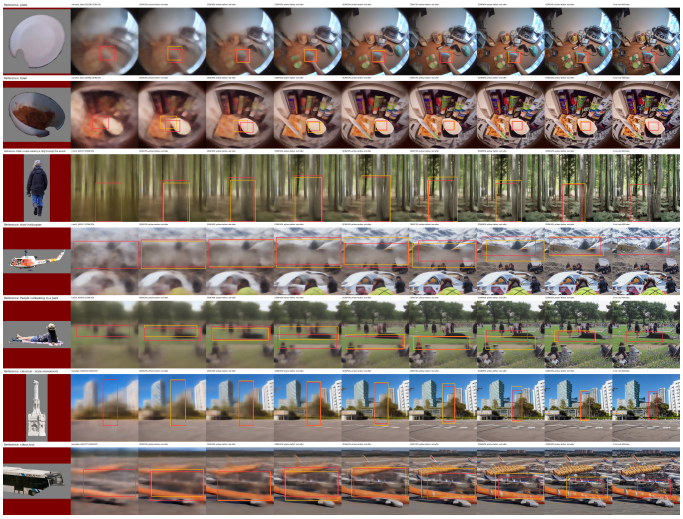
Quantitative Evaluation. As shown in Tab. 1 Coplace+AnyDoor, IP-Adapter and PlanGen+AnyDoor achieve substantial subject-consistency scores than Lay-Your-Scene and PlanGen, demonstrating their stronger capability of subject customization. Coplace+AnyDoor achieves the overall image-quality performance by a small margin, showing competitive results compared with large-scale models. PlanGen+AnyDoor gains an exceptionally low FID score on HD-EPIC. We hypothesize that high-capacity PlanGen adapted particularly well to HD-EPIC that is collected basically in the kitchen scenario.

As shown in Tab. 2. CoPlace achieves better overall performance on metrics evaluating textual alignment (PickScore, ClipScore), diversity (InceptionScore) and subject exclusion (OverlayScore, LeakageScore), while fine-tuned PlanGen consistently yields better FID. We attribute PlanGen model advantage to its larger representational capacity over SD1.5, yet, this advantage reverses in the final image (Tab. 1) when the reference subject is placed without occlusion into CoPlace's empty scene.

Qualitative Evaluation. Fig. 5a exhibits experimental results from CoPlace and baselines. Customized models such as IP-Adapter reproduces the reference subject ignoring text prompts. Lay-Your-Scene and PlanGen sometimes can generate plausible images (line 5), but the subjects fail to maintain consistent identities. In contrast, attributed to the proposed Reciprocal Evolution mechanism, as shown in Fig. 5b, CoPlace enables progressive and collaborative calibration of subject



(a) Lines 1-3, 4-6, 7-9 exhibit results from HiCo-7K, LayoutSAM and HD-EPIC respectively.



(b) Demonstration of scene–bounding-box spiral evolution. Red and yellow boxes denote subject placeholders in the current and previous steps, respectively.

Figure 5: Qualitative results and visualization of scene-bbox spiral evolution. Please zoom in for more details.

placeholder and empty-scene refinement. This reciprocal process allows CoPlace to accurately target subject placement with appropriate geometric properties such as spatial location (line 6 in Fig. 5a) and occlusion relationship (line 3 in Fig. 5a).

Human Evaluation. Five well-educated undergraduates each compared 300 random samples and selected the best among CoPlace+AnyDoor, PlanGen+AnyDoor, and Lay-Your-Scene on subject fidelity, text alignment, and subject-scene harmony. The proposed CoPlace exhibits highest selection rates across three criteria.

Ablation Test. In Tab. 2, CoPlace-SG represents CoPlace model with space gating mechanism deactivated. CoPlace-SG-SE additionally silences the spiral evolution mechanism. CoPlace exhibits similar performance compared with ablation models in terms of image quality, while achieving obviously superior results on OverlayScore and LeakageScore, demonstrating the effectiveness of proposed mechanism.

4.3 Visual Procedural Planning

By disentangling subject editing, CoPlace can also be conveniently adapted to the MPP scenario through our proposed TPP framework. We summarize the procedure planning into three stages, including ingredient preparation, subject processing and post-processing. Fig. 4 demonstrates two procedural image-generation examples. For example, unlike TIP (Lu et al., 2024) and MMP (Tabassum et al., 2025) that generate images solely from text descriptions without considering visual consistency, TPP naturally converts the raw fillets to identical cooked fillets, then find a placeholder for them in a fillets-free scene and finally place the fillets into the scene.

5 Conclusion

We propose TPP, a multi-stage framework for multi-modal procedure planning that addresses the challenges of subject consistency, layout controllability, and subject-background harmony in sequential image generation. By decomposing the generation process into transformation, planning, and placement stages, TPP enables coherent visual procedure generation. Specifically, the proposed CoPlace jointly optimizes subject layout and empty-scene generation, while the space-gating mechanism improves subject placement by reducing semantic interference from textual conditions.

6 FINAL INSTRUCTION

AI USE STATEMENT

In this work, we used generative AI tools for retrieving related work, executing experiments, drafting parts of figures or tables and polishing writing. We have not used generative AI tools for directly drafting sections of this paper, generating synthetic data. Proving mathematical claims are not applicable to this work. We have reviewed all AI-assisted work. We have reviewed all retrieved literature in the early stage. We checked the generated code carefully when conducting experiments. We revised our lines in this paper by ourselves considering the advice from AI. We preserve necessary personal expression rather than directly copying AI work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

See the ICLR 2027 AI Policy for Authors for more details. This statement should not be more than 1 page.

REPRODUCIBILITY STATEMENT

We are committed to improving the reproducibility of this work. We will make our best effort to organize and release the training data used in this study with preprocessing pipelines, and the code for the main experiments. Upon acceptance, these resources will be made publicly available on GitHub to facilitate reproduction and further research by the community.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Saaket Agashe, Jiuzhou Han, Shuyu Gan, Jiachen Yang, Ang Li, and Xin Wang. Agent s: An open agentic framework that uses computers like a human. In *International Conference on Learning Representations*, volume 2025, pp. 22924–22946, 2025.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Stefan Andreas Baumann, Felix Krause, Michael Neumayr, Nick Stracke, Melvin Sevi, Vincent Tao Hu, and Björn Ommer. Continuous, subject-specific attribute control in t2i models by identifying semantic directions. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13231–13241. IEEE, 2025.

Pu Cao, Feng Zhou, Qing Song, and Lu Yang. Controllable generation with text-to-image diffusion models: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6593–6602. IEEE, 2024.

Bo Cheng, Yuhang Ma, Liebucha Wu, Shanyuan Liu, Ao Ma, Xiaoyu Wu, Dawei Leng, and Yuhui Yin. Hico: Hierarchical controllable diffusion model for layout-to-image generation. *Advances in neural information processing systems*, 37:128886–128910, 2024.

Bartosz Cywiński and Kamil Deja. Saeuron: Interpretable concept unlearning in diffusion models with sparse autoencoders. In *Forty-second International Conference on Machine Learning*, 2025.

Weixi Feng, Wanrong Zhu, Tsu-jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. Layoutgpt: Compositional visual planning and generation with large language models. *Advances in neural information processing systems*, 36: 18225–18250, 2023.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

Rohit Gandikota, Joanna Materzyńska, Tingrui Zhou, Antonio Torralba, and David Bau. Concept sliders: Lora adaptors for precise control in diffusion models. In *European Conference on Computer Vision*, pp. 172–188. Springer, 2024.

Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, 2015.

René Haas, Inbar Huberman-Spiegelglas, Rotem Mulayoff, Stella Graßhof, Sami S Brandt, and Tomer Michaeli. Discovering interpretable directions in the semantic latent space of diffusion models. In *2024 ieee 18th international conference on automatic face and gesture recognition (fg)*, pp. 1–9. IEEE, 2024.

Runze He, Bo Cheng, Yuhang Ma, Qingxiang Jia, Shanyuan Liu, Ao Ma, Xiaoyu Wu, Liebucha Wu, Dawei Leng, and Yuhui Yin. Plangen: Towards unified layout planning and image generation in auto-regressive vision language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18143–18154, 2025.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Muhammet Furkan Ilaslan, Ali Köksal, Kevin Qinghong Lin, Burak Satar, Mike Zheng Shou, and Qianli Xu. Vg-tvp: Multimodal procedural planning via visually grounded text-video prompting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3886–3894, 2025.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663, 2023.

Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. *arXiv preprint arXiv:2210.10960*, 2022.

Bingnan Li, Chen-Yu Wang, Haiyang Xu, Xiang Zhang, Ethan Armand, Divyansh Srivastava, Shan Xiaojun, Zeyuan Chen, Jianwen Xie, and Zhuowen Tu. Overlaybench: A benchmark for layout-to-image generation with dense overlaps. *Advances in Neural Information Processing Systems*, 38, 2026.

Dongxu Li, Junnan Li, and Steven Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems*, 36:30146–30166, 2023a.

Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22511–22521. IEEE, 2023b.

Junhao Liu, Jian-Wei Zhang, Tao Huang, Miles Yang, Zhao Zhong, and Liefeng Bo. Think, plan, paint: Layout-aware reasoning for controllable image generation in unified models. *arXiv preprint arXiv:2607.16409*, 2026.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024.

Yujie Lu, Pan Lu, Zhiyu Chen, Wanrong Zhu, Xin Eric Wang, and William Yang Wang. Multimodal procedural planning via dual text-image prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 10931–10954, 2024.

Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. Gaia: a benchmark for general ai assistants. In *International Conference on Learning Representations*, volume 2024, pp. 9025–9049, 2024.

Yong-Hyun Park, Mingi Kwon, Junghyo Jo, and Youngjung Uh. Unsupervised discovery of semantic latent directions in diffusion models. *arXiv preprint arXiv:2302.12469*, 2023.

Toby Perrett, Ahmad Darkhalil, Saptarshi Sinha, Omar Emara, Sam Pollard, Kranti Kumar Parida, Kaiting Liu, Prajwal Gatti, Siddhant Bansal, Kevin Flanagan, et al. Hd-epic: A highly-detailed egocentric video dataset. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23901–23913. IEEE, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22500–22510. IEEE, 2023.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Divyansh Srivastava, Xiang Zhang, He Wen, Chenru Wen, and Zhuowen Tu. Lay-your-scene: Natural scene layout generation with diffusion transformers. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17909–17919. IEEE, 2025.

Afrina Tabassum, Bin Guo, Xiyao Ma, Hoda Eldardiry, and Ismini Lourentzou. Mmplanner: Zero-shot multimodal procedural planning with chain-of-thought object state reasoning. *arXiv preprint arXiv:2509.21662*, 2025.

Kun Wang, Li Yiming, Mingcheng Qu, Aqiang Zhang, Guang Yang, and Tonghua Su. Gala: Hypergraph-guided visual language models for procedural planning. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 19590–19603, 2026.

Li Wang, Boyan Gao, Yanran Li, Zhao Wang, Xiaosong Yang, David A Clifton, and Jun Xiao. Exploring the latent space of diffusion models directly through singular value decomposition. *arXiv preprint arXiv:2502.02225*, 2025.

Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15897–15907. IEEE, 2023.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh J Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, et al. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *Advances in Neural Information Processing Systems*, 37:52040–52094, 2024.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

Hui Zhang, Dexiang Hong, Yitong Wang, Jie Shao, Xinglong Wu, Zuxuan Wu, and Yu-Gang Jiang. Creatilayout: Siamese multimodal diffusion transformer for creative layout-to-image generation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18487–18497. IEEE, 2025a.

Xulu Zhang, Xiaoyong Wei, Wentao Hu, Jinlin Wu, Jiaxin Wu, Wengyu Zhang, Zhaoxiang Zhang, Zhen Lei, and Qing Li. A survey on personalized content synthesis with diffusion models. *Machine Intelligence Research*, 22(5):817–848, 2025b.

Guangcong Zheng, Xianpan Zhou, Xuewei Li, Zhongang Qi, Ying Shan, and Xi Li. Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22490–22499. IEEE, 2023.

A Appendix

A.1 Preliminary

Stable Diffusion (Rombach et al., 2022) performs the diffusion process in a compressed latent space. Given an image x_0, an encoder $E(\cdot)$ first maps it into a latent representation:

$$z_0 = E(x_0). \tag{16}$$

The forward diffusion process gradually adds Gaussian noise to the latent representation:

$$z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \tag{17}$$

where $\epsilon \sim \mathcal{N}(0, I)$ and $\bar{\alpha}_t = \prod_{i=1}^{t}(1 - \beta_i)$ denotes the cumulative noise schedule.

A UNet-based denoising network ϵ_θ is trained to predict the added noise conditioned on timestep t and optional conditions c (e.g., text embeddings):

$$\mathcal{L} = \mathbb{E}_{z_0,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(z_t, t, c)\|_2^2\right]. \tag{18}$$

During inference, the UNet iteratively removes noise from z_t to obtain the clean latent representation, which is finally decoded into an image using the decoder $D(\cdot)$:

$$x_0 = D(z_0). \tag{19}$$

Denoising Diffusion Implicit Models (DDIM) (Song et al., 2020) introduce a deterministic and accelerated sampling strategy for diffusion models. Given the predicted noise $\epsilon_\theta(z_t, t)$, the clean latent state can be estimated as:

$$\hat{z}_0 = \frac{z_t - \sqrt{1 - \bar{\alpha}_t}\epsilon_\theta(z_t, t)}{\sqrt{\bar{\alpha}_t}}. \tag{20}$$

The previous latent state is then obtained by:

$$z_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\hat{z}_0 + \sqrt{1 - \bar{\alpha}_{t-1}}\epsilon_\theta(z_t, t, c). \tag{21}$$

Compared with the original DDPM sampling process, DDIM enables high-quality generation with fewer sampling steps and provides a deterministic denoising trajectory, allowing intermediate latent states and UNet representations to be explicitly accessed during inference.