

Multi-view Image Diffusion via Coordinate Noise and Fourier Attention

Justin Theiss Norman Müller Daeil Kim Aayush Prakash
 Meta Reality Labs

Abstract

Recently, text-to-image generation with diffusion models has made significant advancements in both higher fidelity and generalization capabilities compared to previous baselines. However, generating holistic multi-view consistent images from prompts still remains an important and challenging task. To address this challenge, we propose a diffusion process that attends to time-dependent spatial frequencies of features with a novel attention mechanism as well as novel noise initialization technique and cross-attention loss. This Fourier-based attention block focuses on features from non-overlapping regions of the generated scene in order to better align the global appearance. Our noise initialization technique incorporates shared noise and low spatial frequency information derived from pixel coordinates and depth maps to induce noise correlations across views. The cross-attention loss further aligns features sharing the same prompt across the scene. Our technique improves SOTA on several quantitative metrics with qualitatively better results when compared to other state-of-the-art approaches for multi-view consistency.

1. Introduction

In recent years, significant breakthroughs have been made in text-conditional image generation [14, 19, 21, 23, 30]. However, when extending single-view image generation to multi-view and video generation from text prompts [1, 6–8, 11, 12, 17, 20, 24, 27] there remains considerable challenges, particularly around the consistency of a scene’s geometry and appearance. To this end, recent works [1, 2, 24] implement attention modules that process all views simultaneously [2, 24]. This aims to align features across views by incorporating cross-attention modules into the standard diffusion model architecture. Moreover, MVDiffusion [24] uses known camera pose and depth information in order to find corresponding points for attention across different views. Similarly, ConsistI2V [20] proposed changes to cross attention between views, such as attending to the local neighborhood around a query index for each view, but performance seems constrained to video sequences with high

temporal sampling. In the case of more general multi-view image generation (e.g., panoramas), such a method may not adequately handle larger changes in camera pose between views. Specifically, for methods relying on high overlap between frames, the appearance in areas with less overlap across the scene often exhibit stark changes (see Figure 1). Improving consistency in non-overlapping regions is therefore important for ensuring consistency in the global appearance.

Another exciting direction to improve the multi-view consistency of text-to-image is through the role of noise initialization by using shared noise [6], correlated noise [17], or low spatial frequency components of images [27]. These studies have shown that overall appearance can be improved by combining shared and independent components when initializing noise for multi-view generation. One possible explanation for this effect is the recently observed gap between noise used during training and inference [13]. At the noisiest time step during training, low spatial frequency information regarding the image is still present; however, during inference, this information is missing when sampling from Gaussian noise. In this work, we leverage this initialization gap in order to improve consistency across generated images by inducing low-frequency correlations across noise samples without requiring access to ground truth images [20] or costly sampling steps to generate a starting layout [27].

We address the challenge of multi-view consistent text-to-image generation through a novel diffusion-based method that combines noise initialization with Fourier-based attention to guide image generation toward a consistent appearance. Building on recent work highlighting the gap in signal-to-noise ratio between noise samples used during training and inference, we propose a method for coordinate-based noise initialization that induces low spatial frequency correlations in the noise samples across views. We further propose an attention module that aligns non-overlapping regions across views by attending to progressively higher spatial frequency features across denoising time steps. Finally, we introduce a prompt-based cross-attention loss that ensures attention between prompt tokens and each view is consistent with the ground truth

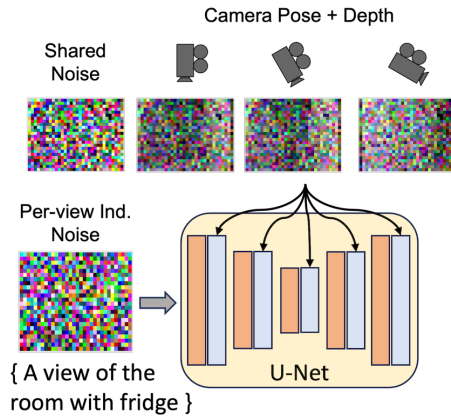
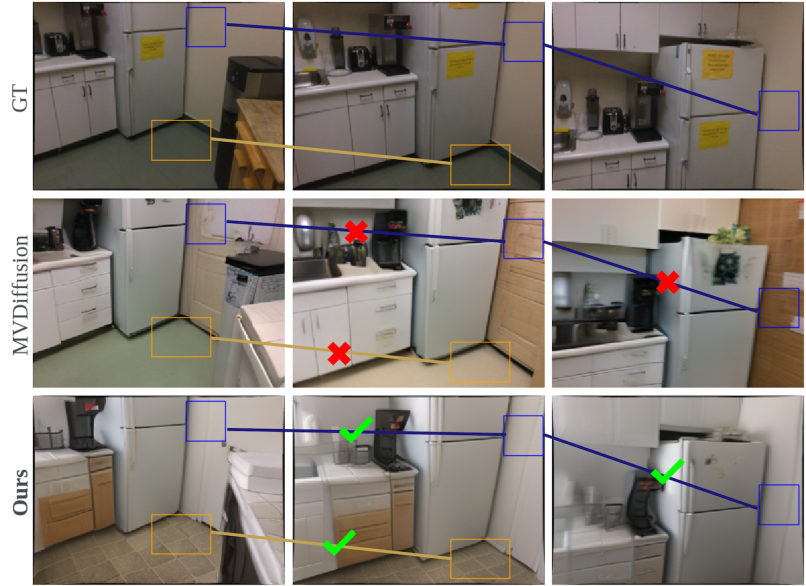


Figure 1. We propose a method that addresses the lack of consistency in multi-view image generation by aligning appearance in non-overlapping regions of multi-view scenes (*left*). Compared to MVDiffusion, our approach improves the consistency of textures and geometry, particularly in non-overlapping regions (*e.g.*, floor and walls; *right*).

scene. Overall, our method improves multi-view consistency by initializing with pose-dependent noise, attending to frequency-dependent features in non-overlapping regions, and ensuring consistent semantic relationships via a prompt-based cross-attention loss.

In order to understand the effect of various design choices for more disparate views, we evaluate performance for multi-view consistency using two settings: panoramic and depth-conditioned image generation. We demonstrate quantitatively and qualitatively that these design choices improve image quality and multi-view consistency over state-of-the-art approaches. We summarize our contributions as follows:

- We introduce a novel noise initialization technique, which incorporates shared noise and low spatial frequency information across views without time-prohibitive diffusion inversion or access to real images.
- We introduce a novel attention mechanism (Fourier-based Attention) that attends to shared-noise features across non-overlapping regions of a scene.
- Finally, we introduce a novel cross-attention loss that aligns multi-view prompt cross-attention maps with the ground truth attention maps, improving alignment of features sharing the same prompt across views.

2. Related Work

2.1. Text-to-Image Diffusion Models

The field of text-to-image diffusion models has undergone considerable advancements, with significant contributions from models like DALL-E 2 [19], GLIDE [15], Latent Diffusion Models (LDMs) [21], and Imagen [23]. These models excel in generating photo-realistic images from text prompts, combining the efficiency of large-scale diffusion models with the sophistication of pre-trained language models. Additional control over the image output is possible through manipulation of model cross-attention layers, as demonstrated in Prompt-to-Prompt [9], Attend-and-Excite [4], and FreestyleNet [28]. This cross-attention control has also led to improvements in multi-view consistency. Our model extends the single-view text-to-image LDM into the multi-view domain with additional conditioning on camera pose and/or depth.

2.2. Multi-view Consistency in Image Diffusion Models

The pursuit of multi-view consistency in image generation has led to several noteworthy advancements. MultiDiffusion [1] focused on fusing diffusion paths for controlled image generation, addressing the seamless integration of multiple views. SyncDiffusion [12] synchronizes joint diffusions for coherent montage creation, using gradient descent from perceptual similarity loss to align multiple diffusions. DiffCollage [31] generates large content

by merging results from overlapping nodes represented by a factor graph. However, all these techniques either do not generate true panorama (left and right corners mismatch) or have visible artifacts in the generated images. TokenFlow [7] generates multi-view consistent edits in videos by propagating features based on inter-frame correspondence, but needs access to ground truth noise samples. MVDiffusion [24] improved multi-view image generation by embedding correspondence-aware attention in diffusion models, optimizing for consistency across multiple views. However, MVDiffusion struggles to generate multi-view consistent images in non-overlapping regions. Our method explicitly attends to such regions to align appearance.

2.3. Noise Initialization in Diffusion Models

Recent works have investigated the gap in noise initialization between training and inference for diffusion models [13, 29], noting that there is an information leak that occurs even at highest noise levels during training. Other works have leveraged this information leak as a way to improve consistency in appearance by incorporating low spatial frequency information [20, 27] or the mean of images from a given class [29]. In video diffusion models, initializing with a weighted combination of shared and independent noise across frames [6] or inducing long-range correlations via noise rescheduling [17] have similarly achieved greater global consistency. Our method instead leverages 3D coordinate information to inform *spatial* structure of the scene.

3. Method

The method section is organized as follows: we first cover the preliminaries regarding diffusion models, then we propose our method for noise initialization, Fourier-based attention, and prompt-based cross-attention loss. We end the section with a description of the full training paradigm.

3.1. Preliminary

Image diffusion models are trained to model a data distribution p_{data} by iteratively denoising an image $\mathbf{x}$ from a random Gaussian noise sample across a sequence of T time steps. Latent Diffusion Models (LDMs) instead operate on a latent representation $\mathbf{z}$ from a pre-trained VAE autoencoder (*i.e.*, $\mathbf{z} := \mathcal{E}(\mathbf{x})$ and $\mathbf{x} := \mathcal{D}(\mathbf{z})$). During the forward diffusion process, noise is added to the latent $\mathbf{z}_0$ at each time step t :

$$q(\mathbf{z}_t|\mathbf{z}_{t-1}) = \mathcal{N}(\mathbf{z}_t; \sqrt{\alpha_t}\mathbf{z}_{t-1}, \sigma_t^2\mathbf{I}), \quad (1)$$

where the noise schedule is defined by parameters α_t and σ_t^2 derived from a predefined variance schedule $\beta_1, \dots, \beta_T$ with $\sigma_t^2 = \beta_t$ and $\alpha_t = 1 - \beta_t$. In practice, the forward diffusion can be determined in a single step:

$$q(\mathbf{z}_t|\mathbf{z}_0) = \mathcal{N}(\mathbf{z}_t; \sqrt{\bar{\alpha}_t}\mathbf{z}_0, (1 - \bar{\alpha}_t)\mathbf{I}) \quad (2)$$

where $\bar{\alpha}_t = \prod_{i=1}^t \alpha_i$. Therefore, the noisy latent $\mathbf{z}_t$ can be directly sampled with Gaussian noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$:

$$\mathbf{z}_t = \sqrt{\bar{\alpha}_t}\mathbf{z}_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon. \quad (3)$$

The LDM is then trained to approximate the reverse diffusion process, in order to obtain the latent $\mathbf{z}_0$ from Gaussian noise $\mathbf{z}_T$:

$$p_\theta(\mathbf{z}_{0:T}) = p(\mathbf{z}_T) \prod_{t=1}^T p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t) \quad (4)$$

$$p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t) = \mathcal{N}(\mathbf{z}_{t-1}; \mu_\theta(\mathbf{z}_t, t), \Sigma_\theta(\mathbf{z}_t, t)), \quad (5)$$

where μ_θ and Σ_θ are predicted by the denoising network ϵ_θ , typically implemented using the U-Net [22] architecture. The denoising network ϵ_θ is then trained to predict the ground truth noise ϵ by minimizing the following objective function:

$$\mathcal{L}_{LDM} = \mathbb{E}_{t, \mathbf{z}_0, \epsilon} \|\epsilon - \epsilon_\theta(\mathbf{z}_t, t)\|_2^2. \quad (6)$$

3.2. Coordinate-based Noise Initialization

During inference, the noisy latent $\mathbf{z}_T$ is sampled from $\mathcal{N}(0, \mathbf{I})$ and the denoised sample $\mathbf{z}_0$ is decoded using the pre-trained decoder $\mathcal{D}$ to obtain the generated image $\hat{\mathbf{x}}$. However, as noted in recent works [13, 29], there is an SNR gap between the noisy latent $\mathbf{z}_T$ used during training and inference that leads to a information leak provided to the model during training but not during inference. This suggests that consistency in multi-view image generation may be improved by incorporating shared noise and/or low spatial frequencies from a target image when initializing the noise latent $\mathbf{z}_T$. Indeed, rather than using the latent sampled from $\mathcal{N}(0, \mathbf{I})$, previous works have observed that incorporating latent features from real images [20, 27] into $\mathbf{z}_T$ can improve the consistency when generating multiple images (*e.g.*, a video sequence). In our experiments, we explore the use of noise initialization methods that do not require access to the diffusion-inverted latent features $\mathbf{z}_0$ of real images during inference.

Based on the aforementioned gap between the latent $\mathbf{z}_T$ used during training and inference, it is intuitive to initialize noise using Equation 3 by replacing the original latent $\mathbf{z}_0$ with some shared noise or image features ϵ_{shared} :

$$\hat{\mathbf{z}}_T^i = \sqrt{\bar{\alpha}_T}\epsilon_{\text{shared}} + \sqrt{1 - \bar{\alpha}_T}\epsilon^i, \quad (7)$$

where ϵ_{shared} is shared across all views and $\epsilon^i \sim \mathcal{N}(0, \mathbf{I})$ is sampled independently per view. Setting ϵ_{shared} to random Gaussian noise has the effect of providing similar shared

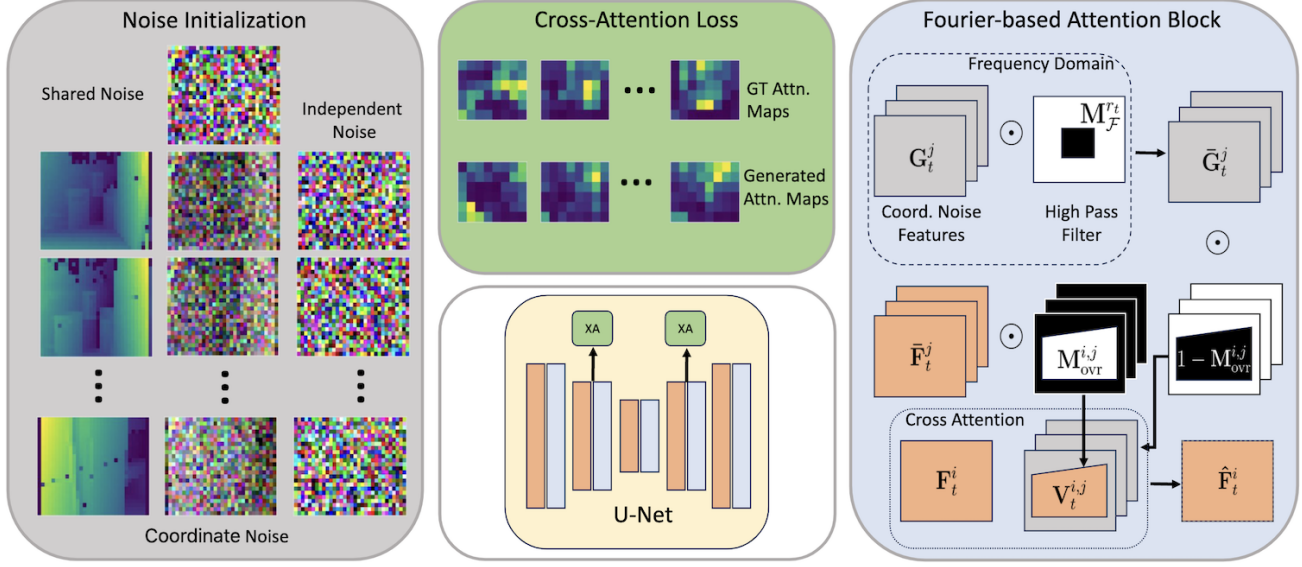


Figure 2. Overview of our proposed technique. (Left) We initialize noise by sampling Gaussian noise shared across views as well as independent per view. We combined the shared noise with depth or transformed pixel coordinates to obtain *Coordinate Noise* (Sec. 3.2), which provides a low spatial frequency bias to inform the overall structure of the scene. (Center) We add our Fourier-based Attention (FBA) blocks (blue, see right panel) within the U-Net architecture and introduce a cross-attention loss (Sec. 3.4) to ensure consistent spatial relationships across views. (Right) Finally, our novel attention module (Sec. 3.3) time-dependent spatial frequencies of features generated from the *Coordinate Noise* in non-overlapping regions to better align the global appearance across the scene.

statistics across views, but does not necessarily provide information about the scene generally.

In order to provide information about the scene, we propose incorporating depth maps (where available) and transformed pixel coordinates based on camera pose information between each view and a reference (e.g., center view of a panorama). By utilizing the camera pose and (optionally) depth information of the scene, we are able to induce low spatial frequency correlations implicitly across the scene. We then combine the coordinate-based component with the shared noise to obtain our noise initialization:

$$\hat{\epsilon}^i = w * \mathbf{c}^i + (1 - w) * \epsilon_{\text{shared}}, \quad (8)$$

$$\hat{\mathbf{z}}_T^i = \sqrt{\alpha_T} \hat{\epsilon}^i + \sqrt{1 - \alpha_T} \epsilon^i, \quad (9)$$

where $\mathbf{c}^i$ represents the low-frequency coordinate/depth-map information for view i , which is linearly combined with the shared noise ϵ_{shared} using weight w . For depth conditioning, we set $\mathbf{c}^i$ as the normalized depth maps per view, whereas for panoramic images we use the normalized coordinates for each view transformed into the coordinate space of the center view (see supplemental material for further detail). We refer to this combination of coordinate-based information and shared noise as *coordinate noise*. When $w = 0$, there is no low-frequency condition – a setting we refer to as *simple shared noise*. The updated latent $\hat{\mathbf{z}}_T$ is then used in place of $\mathbf{z}_T$ when initializing noise for image

generation.

3.3. Fourier-based Attention Module

Based on the insights from works [6, 20, 27] incorporating shared noise or low spatial frequency information when generating multi-view images, we hypothesize that attending to coordinate noise features – particularly in non-overlapping regions of multi-view images – will improve the consistency in global appearance across a generated scene. Building on MVDiffusion’s [24] *correspondence-aware attention* (CAA) modules, we propose a Fourier-based attention (FBA) module that incorporates coordinate noise features with spatial frequencies selected dependent on the denoising time step. Similar to the CAA blocks, each FBA block contains the attention module and a residual network with zero-initialized convolution layers. Whereas the CAA modules are intended to attend to corresponding points in overlapping image regions, the intuition of our approach is to inject shared noise to better align the overall scene appearance in the non-overlapping image regions.

Specifically, for a given time step t and noise initialization (i.e. Equation 9), let $\mathbf{F}_t^i$ be the feature maps of the U-Net denoising network for the view indexed by $i \in [0, N - 1]$. Let $\mathbf{G}_t^j$ then be the features obtained when using the *coordinate noise* $\hat{\epsilon}$ (i.e. Equation 8) to set $\mathbf{z}_t$ in Equation 3. Since these feature maps are the target of the attention module, we gather $\mathbf{G}_t^j$ from each preceding layer

without applying the FBA blocks. We then apply the Fast Fourier Transform (FFT; Equation 10) to obtain the spatial frequency domain representations of features $\mathbf{G}_t^i$, allowing us to select and attend to higher-frequency features to align appearance, where lower frequencies more represent image structure [27]. The inverse transform $\mathcal{F}^{-1}$ is similarly applied to transform back to the image domain (*i.e.* FFT and its inverse are applied to the height and width dimensions).

$$\mathcal{F}(m, n) = \sum_{h, w} \mathbf{x}(h, w) \exp -j2\pi \left(\frac{h}{H}m + \frac{w}{W}n \right), \quad (10)$$

$$j^2 = -1$$

To implement this method, we combine the *correspondence-aware attention* within overlapping regions and our *Fourier-based attention* in non-overlapping regions. Using the known homography matrices relating each view, we can obtain the mask of overlapping regions $\mathbf{M}_{\text{ovr}}^{i,j}$ between views i and j . Formally, for the set of source feature maps $\mathbf{F}_t^i$, we select the corresponding features $\bar{\mathbf{F}}_t^j$ in overlapping regions of each target view j and the spatially-filtered features $\bar{\mathbf{G}}_t^j$ in non-overlapping regions. Following [24], the features in overlapping regions of target views are interpolated from the target coordinate space v to map onto the corresponding locations in the source coordinate space u (see [24] for further detail):

$$\bar{\mathbf{F}}_t^j(v^j) = \mathbf{F}_t^j(v^j) + \gamma(u_*^j - u), \quad (11)$$

where $\gamma(\cdot)$ represents the positional encoding of the displacement between corresponding coordinates u_*^j and original coordinates u in the source view.

The *coordinate noise* features $\mathbf{G}_t^j$ are spatially filtered by masking portions of the frequency spectrum dependent on the time step. Specifically, for each time step we select spatial frequencies proportional to the noise level by creating a mask with ones everywhere except within a central region whose radius is dependent on the time step:

$$r_t = 1 - \frac{t}{T}, \quad (12)$$

$$\mathbf{M}_{\mathcal{F}}^{r_t} = (1 - \mathbb{1}_{(h,w) \in [-r_t H : r_t H, -r_t W : r_t W]}), \quad (13)$$

where H and W are the height and width of the U-Net features, respectively. The mask $\mathbf{M}_{\mathcal{F}}^{r_t}$ therefore represents the spatial frequencies corresponding to the *coordinate noise* features $\mathbf{G}_t^j$. The mask is then element-wise multiplied with the frequency signal to obtain the non-overlapping features to be attended, as shown in Equation 14 below.

$$\bar{\mathbf{G}}_t^j = \mathcal{F}^{-1}(\mathbf{M}_{\mathcal{F}}^{r_t} \odot \mathcal{F}(\mathbf{G}_t^j)) + \gamma(1 - r_t), \quad (14)$$

where $\gamma(1 - r_t)$ represents the positional encoding of the time step-dependent radius. Finally, the target features of

overlapping and non-overlapping regions are combined using the mask of overlapping regions $\mathbf{M}_{\text{ovr}}^{i,j}$:

$$\mathbf{V}_t^{i,j} = \mathbf{M}_{\text{ovr}}^{i,j} \odot \bar{\mathbf{F}}_t^j + (1 - \mathbf{M}_{\text{ovr}}^{i,j}) \odot \bar{\mathbf{G}}_t^j. \quad (15)$$

We then apply attention from each query source view $\mathbf{F}^i$ to the set of target views $\mathbf{V}^i$:

$$\hat{\mathbf{F}}^i = \text{SoftMax}([\mathbf{W}_Q \mathbf{F}^i] \cdot [\mathbf{W}_K \mathbf{V}^i]) \mathbf{W}_V \mathbf{V}^i, \quad (16)$$

where $\mathbf{W}_Q$, $\mathbf{W}_K$, and $\mathbf{W}_V$ are the weights corresponding to the query, key, and value commonly used in attention modules [25].

3.4. Prompt Cross Attention Loss

In order to improve the spatial consistency of features across views, we propose a novel loss that ensures that the cross attention maps between each prompt and view are consistent with those in the ground truth scene. This method takes inspiration from the cross attention loss proposed in [16], which was shown to improve structural consistency during image-to-image editing with diffusion models. We extend this cross attention loss to multi-view image generation by computing it on the attention between each view's prompt and all other view features. This ensures that the prompt-based attention between disparate views is consistent with the ground truth scene.

We implement the cross attention loss at each 16×16 resolution cross attention module. In order to obtain the ground truth attention maps $\mathcal{M}_t^l$, we first pass the clean latent views $\mathbf{z}_0^{1:n}$ through the U-Net to collect the noise-free attention maps. The cross attention loss is then computed at each applicable layer l as follows:

$$\mathcal{L}_{XA}^l = \|\mathcal{M}_t^l - \mathcal{M}_0^l\|. \quad (17)$$

3.5. Training Paradigm

In order to train our FBA blocks, we start from a diffusion model trained on single views. For depth-to-image training, the diffusion model is first fine-tuned to generate images at the 192×256 image resolution. This training is done using single-view images only. During training of the FBA blocks, we randomly select a sequence of n partially overlapping views from the dataset and a single time step for all views $t \sim \mathcal{U}[1, T]$. During this stage, we keep the original U-Net model parameters frozen and train the proposed FBA blocks end-to-end to minimize the following overall loss function:

$$\mathcal{L} = \mathcal{L}_{LDM} + \lambda \sum_{l \in L} \mathcal{L}_{XA}^l, \quad (18)$$

where L denotes the set of layer indices corresponding to attention maps processing 16×16 spatial resolution and λ is set to 10.

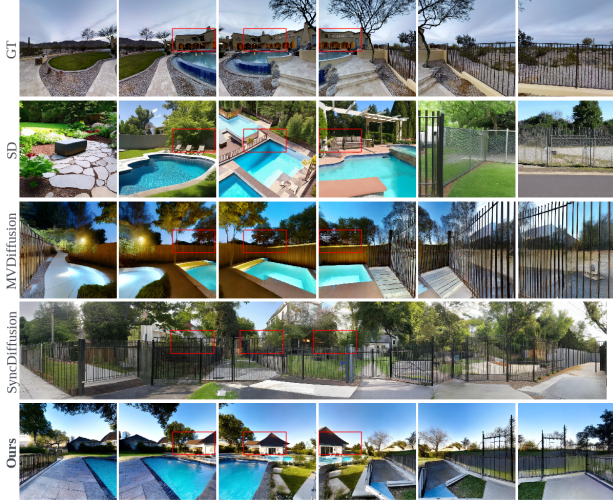


Figure 3. Qualitative comparison for panoramic image generation. Colored boxes highlight misalignment with prompt “a house with a pool in the backyard”. See Section 4.4 for further detail.

Method	FID ↓	CLIP Score ↑
Baseline LDM [21]	45.6	25.6
MVDiffusion [24]	30.3	24.3
SyncDiffusion [12]	51.4	20.0
Ours	22.36	24.7

Table 1. Quantitative evaluation of image quality in panoramic experiments compared with baseline.

Method	PSNR ↑	Ratio ↑	I-LPIPS ↓
GT	37.7	1.00	0.71
Baseline LDM [21]	9.10	0.24	0.80
MVDiffusion [24]	22.2	0.60	0.79
SyncDiffusion [12]	-	-	0.64
Ours	24.7	0.66	0.75

Table 2. Quantitative evaluation of multi-view consistency. Bold text indicates best performance among methods addressing multi-view consistency.

4. Experiments

We evaluate our method in two settings: multi-view panoramic and depth-conditioned image generation. To evaluate panoramic image generation, we use the Matterport3D¹ dataset [3] consisting of 10,912 panoramic indoor scenes. Following [24], we separate the dataset into 9820 panoramic sequences used for training and 1092 for evaluation. To evaluate performance of multi-view depth-to-image generation, we use the ScanNet dataset [5], containing 1513 scenes for training and 100 scenes for evaluation. We discuss method implementation details in Section 4.1,

¹<https://matterport.com/legal/matterport-end-user-license-agreement-academic-use-model-data>

baselines in Section 4.2, quantitative, qualitative and ablation results in Section 4.3, Section 4.4 and Section 4.5 respectively. Additional results and implementation details can be found in the Supplementary Materials.

4.1. Implementation Details

We implement our method with PyTorch using the latent diffusion model architecture provided from Diffusers [26]. During training, we freeze the parameters of the denoising U-Net and train only our newly added modules. We train each method using 4 nodes with 8 x A100 GPUs each for 20 epochs in the depth-to-image experiment and 10 epochs in the panoramic experiment. We use a per-GPU batch size of 1 and a learning rate $1e^{-4}$ and $2e^{-4}$ for the depth-conditioned and panoramic experiments, respectively.

During inference, we generate 8 views simultaneously for both depth-to-image and panoramic experiments. For panoramic image generation, each view is separated by a rotation angle of 45 degrees. For depth-to-image generation, we follow the method described in [24] for generating key frames and interpolation for denser image generation. The key-frame views are curated to maintain approximately 65% overlap between each pair of key frames. For interpolating between views, the generated key frames are used to condition the generation of the interpolated views as described in [24].

4.1.1 Evaluation Metrics

We utilize multiple metrics to evaluate image quality and multi-view consistency. To evaluate the image quality of generated multi-view scenes, we compute the following metrics:

- Frechet Inception Distance (FID) [10]: measures the distribution gap between generated and real images.
- CLIP Score (CS) [18]: measures the text and generated image similarity using the CLIP model.

In order to evaluate consistency of generated multi-view images, we use the following metrics:

- Overlap Peak Signal-to-Noise Ratio (PSNR) [24]: PSNR between all overlapping regions, compared as a ratio between generated and real images.
- Intra-LPIPS [32]: measures the coherence of panoramic images, computed as the average LPIPS distance of all combinations of generated image pairs for a scene.

We use the same evaluation method for each experiment as described in [24]. In brief, to evaluate multi-view consistency we compute overlapping PSNR ratios between consecutive generated images relative to the ground truth comparisons.



Figure 4. Qualitative comparison for multi-view depth-to-image generation. Colored boxes highlight inconsistencies in baselines relative to our method. Blue, red and white boxes demonstrate how small objects, large objects and environment (*i.e.* non-overlapping regions) resp. change appearance in our baselines. We qualitatively outperform our baselines. See Section 4.4 for further detail.

Method	FID ↓	CLIP Score ↑	Method	PSNR ↑	Ratio ↑	Intra-LPIPS ↓
ControlNet [30]	38.3	20.5	GT	15.4	1.00	0.58
MVDiffusion [24]	23.7	24.3	ControlNet [30]	9.05	0.62	0.70
Ours	27.0	23.8	MVDiffusion [24]	13.0	0.87	0.67
			Ours	13.9	0.94	0.64

Table 3. Quantitative comparison with baseline methods for depth-conditioned experiments. Left: image quality, right: multi-view consistency. Bold text indicates best performance.

4.2. Baselines

We evaluate our performance against the following baseline methods for the panoramic experiment:

- MVDiffusion [24] uses the *correspondence-aware* attention module to attend to a nearby set of views for panoramic experiments.
- Baseline LDM [21] constitutes the baseline pre-trained model upon which MVDiffusion is trained.
- SyncDiffusion [12] is a training-free method designed for panoramic image generation with diffusion models.

To evaluate performance for the depth-to-image experiment, we compare against the following baselines:

- MVDiffusion [24] incorporates a *correspondence-aware* attention module (CAA) that attends to nearby views with corresponding points. As described above, we utilize this CAA mechanism within our own attention module.
- ControlNet [30] is a popular method for conditioning LDMs, which in our experiments can be used for evaluating depth-to-image generation.

Whereas our and other baseline methods generate n views in parallel, SyncDiffusion [12] generates a single panoramic image with 512×3072 resolution using a single text prompt. In order to compare against the other baselines, we first combine the per-view text prompts used in our method into a single prompt describing the full scene. Then following image generation, we split the panoramic image into six non-overlapping views with resolution 512×512 as described in [12]. We use these views for quantitative evaluations of FID, CLIP Score, and Intra-LPIPS.

4.3. Quantitative Evaluation

4.3.1 Panoramic Experiment

We report quantitative results for the panoramic image generation experiment in Tables 1 and 2. As shown in the tables, our method consistently outperforms the baselines across most metrics, particularly FID and overlapping PSNR. Although SyncDiffusion achieves a lower Intra-LPIPS, the value is far lower even than the ground truth images (0.64 vs. 0.71). This is likely due to the fact that their method aims to increase coherence across all views, resulting in similar content repeated across the scene (*e.g.*, see Figure 3). This is also supported by their relatively low CLIP Score (20.0 vs. 24.7 in our method), which indicates

that their generated images do not respect the provided text prompts. Meanwhile, among methods that aim to improve multi-view consistency, ours has the highest CLIP Score.

4.3.2 Depth-to-Image Experiment

We report quantitative results for the depth-to-image experiment in Table 3. As seen in the table, our method greatly improves the multi-view consistency compared to MVDiffusion (0.94 vs. 0.87 ratio and 0.64 vs. 0.67). In addition to non-overlapping improvements from FBA blocks, we attribute improvements in overlapping regions to two primary differences: 1) coordinate-based noise initialization better informs scene structure and 2) the prompt cross attention loss improves prompt-spatial alignment across views. However, we do observe slightly lower performance in the depth-to-image experiment in terms of FID, while CLIP Score demonstrates competitive performance, which we discuss in further detail in the supplemental material.

4.4. Qualitative Evaluation

4.4.1 Panoramic Experiment

We show qualitative results in Figure 3 for the panoramic experiment. When compared with MVDiffusion and SyncDiffusion, we observe several instances where generated views are missing attributes from the provided prompt. In this case, the central views were conditioned on the prompt “a house with a pool in the backyard”, but the generated scene from MVDiffusion’s method does not contain a house. SyncDiffusion’s method generates a house but fails to generate the pool. We hypothesize our method achieves better prompt alignment in such cases due to the XA loss, which trains the model to generate images that preserve the attention maps between each prompt and all views.

4.4.2 Depth-to-Image Experiment

As shown in Figure 4, the other baseline methods exhibit multiple inconsistencies across generated views. Specifically, the blue box in the first three columns demonstrate how small objects may change appearance when generating with SD or MVDiffusion. The red box in the middle columns highlight changes in texture color for larger objects, where MVDiffusion the desk color changes from brown to white. Finally, the white box in columns 2 and 8 showcase our method’s ability to generate images with a globally consistent appearance, whereas in MVDiffusion’s scene the floor texture changes across the scene. We argue that the incorporation of our FBA blocks, which attend to the non-overlapping regions, help in achieving higher consistency across more disparate views of a scene. This is supported by Figure 5, which demonstrates that when us-



Figure 5. Color/texture inconsistencies using CAA (MVDiffusion) vs. FBA (ours) blocks.

Method	FID	CS	PSNR	Ratio	I-LPIPS
Shared Noise	22.1	24.7	23.6	0.63	0.79
Coord. Noise	19.5	24.9	24.2	0.65	0.78
FBA Blocks	21.0	24.9	23.9	0.64	0.78
Full Model	22.4	24.7	24.7	0.66	0.75

Table 4. Ablation of FBA Blocks.

ing MVDiffusion’s CAA blocks vs. our FBA blocks abrupt changes in colors and textures are observed.

4.5. Ablation Studies

In Table 4, we report quantitative results of an ablation study for the panoramic image generation experiment (see supplemental material for qualitative comparisons). We start by evaluating performance when only using *Shared Noise* (Eqn. 7), which greatly improves FID relative to the MVDiffusion baseline, but provides more modest improvements in multi-view consistency. Using instead the *Coordinate Noise* (Eqn. 8) for initialization provides further improvements in both image quality and multi-view consistency. We observe similar performance when introducing the FBA blocks (Sec. 3.3); however, when combined with the cross-attention loss (Sec. 3.4), we observe our best overall performance. Figure 5 further shows the qualitative improvements when using FBA vs. CAA blocks.

5. Conclusion

In this paper we address the challenge of multi-view consistent text-to-image generation. We propose a diffusion model that utilizes the Fourier space to select features for attention in non-overlapping regions. We further propose a novel noise initialization technique and cross-attention that ensure higher multi-view consistency in the overlapping regions. As shown qualitatively and quantitatively we outperform SOTA baselines and achieve multi-view consistency while maintaining the diversity in the generated images. In the future, we want to extend this work to generate high-fidelity, multi-view and temporally-consistent videos from prompts, conditioned on depth-maps.

References

- [1] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. 2023. 1, 2
- [2] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023. 1
- [3] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017. 6
- [4] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4), 2023. 2
- [5] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 6
- [6] Songwei Ge, Seungjun Nah, Guilin Liu, Tyler Poon, Andrew Tao, Bryan Catanzaro, David Jacobs, Jia-Bin Huang, Ming-Yu Liu, and Yogesh Balaji. Preserve your own correlation: A noise prior for video diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22930–22941, 2023. 1, 3, 4
- [7] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373*, 2023. 1, 3
- [8] Jiayi Gu, Shicong Wang, Haoyu Zhao, Tianyi Lu, Xing Zhang, Zuxuan Wu, Songcen Xu, Wei Zhang, Yu-Gang Jiang, and Hang Xu. Reuse and diffuse: Iterative denoising for text-to-video generation. *arXiv preprint arXiv:2309.03549*, 2023. 1
- [9] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. Prompt-to-prompt image editing with cross-attention control. In *The Eleventh International Conference on Learning Representations*, 2022. 2
- [10] 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. 6
- [11] Lukas Höllein, Ang Cao, Andrew Owens, Justin Johnson, and Matthias Nießner. Text2room: Extracting textured 3d meshes from 2d text-to-image models. *arXiv preprint arXiv:2303.11989*, 2023. 1
- [12] Yuseung Lee, Kunho Kim, Hyunjin Kim, and Minhyuk Sung. Syncdiffusion: Coherent montage via synchronized joint diffusions. *arXiv preprint arXiv:2306.05178*, 2023. 1, 2, 6, 7
- [13] Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. Common diffusion noise schedules and sample steps are flawed. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5404–5411, 2024. 1, 3
- [14] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 1
- [15] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*. PMLR, 2022. 2
- [16] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023. 5
- [17] Haonan Qiu, Menghan Xia, Yong Zhang, Yingqing He, Xintao Wang, Ying Shan, and Ziwei Liu. Freenoise: Tuning-free longer video diffusion via noise rescheduling. *arXiv preprint arXiv:2310.15169*, 2023. 1, 3
- [18] 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. 6
- [19] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1, 2
- [20] Weiming Ren, Harry Yang, Ge Zhang, Cong Wei, Xinrun Du, Stephen Huang, and Wenhui Chen. Consisti2v: Enhancing visual consistency for image-to-video generation. *arXiv preprint arXiv:2402.04324*, 2024. 1, 3, 4
- [21] 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. 1, 2, 6, 7
- [22] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 3
- [23] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35, 2022. 1, 2
- [24] Shitao Tang, Fuyang Zhang, Jiacheng Chen, Peng Wang, and Yasutaka Furukawa. Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. *arXiv preprint arXiv:2307.01097*, 2023. 1, 3, 4, 5, 6, 7

- [25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 5
- [26] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. <https://github.com/huggingface/diffusers>, 2022. 6
- [27] Tianxing Wu, Chenyang Si, Yuming Jiang, Ziqi Huang, and Ziwei Liu. Freeinit: Bridging initialization gap in video diffusion models. *arXiv preprint arXiv:2312.07537*, 2023. 1, 3, 4, 5
- [28] Han Xue, Zhiwu Huang, Qianru Sun, Li Song, and Wenjun Zhang. Freestyle layout-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2
- [29] Jeffrey Zhang, Shao-Yu Chang, Kedan Li, and David Forsyth. Preserving image properties through initializations in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5242–5250, 2024. 3
- [30] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 1, 7
- [31] Qinsheng Zhang, Jiaming Song, Xun Huang, Yongxin Chen, and Ming-Yu Liu. Diffcollage: Parallel generation of large content with diffusion models. *arXiv preprint arXiv:2303.17076*, 2023. 2
- [32] 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. 6