Align Images Before You Generate

Shihua Zhang, Qiuhong Shen, Xinchao Wang*
National University of Singapore, Singapore
suhzhang001@gmail.com, qiuhong.shen@u.nus.edu, xinchao@nus.edu.sg

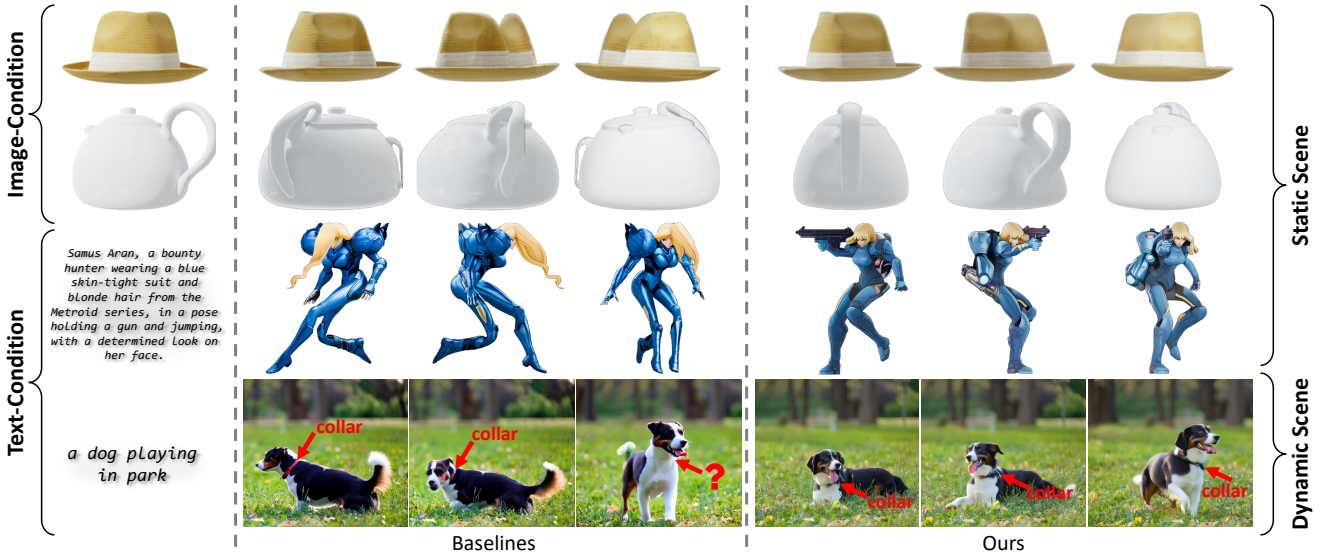

Figure 1. **CorrAdapter** is a versatile spatiotemporal adapter tailored to enhance consistency in multi-image generation by solely mining guidance from diffusion-native correspondence. It is applicable to both dynamic and static scenarios, and supports conditioning on either images or text. CorrAdapter operates alongside baseline generators as a plug-and-play module that aligns the generated outputs. With this adapter employed, generators consistently improve visual coherence and maintain semantic fidelity across generated images.

Abstract

Multi-image diffusion models can generate images like multi-views or videos to describe static or dynamic scenes, yet texture and structure drift persist, severely undermining the spatiotemporal consistency. Addressing this issue remains challenging, especially without any external geometric or semantic priors during the pure generative inference. In this paper, we introduce CorrAdapter, a plug-and-play adapter that discovers and exploits an innate property of the multi-image diffusion itself, aligning all output images before they are in fact generated. Specifically, CorrAdapter designs a bypass branch for transformer blocks in the multi-image diffusion model, encompassing a native correspondence constructor that builds reliable correspondences from the diffusion model's intermediate features, and an aligned area aggregator that integrates messages from only matching regions to avoid ambiguous information interactions. Given the native correspondences as guidance, CorrAdapter can enhance spatiotemporal consistency without any auxiliary inputs, and remains training-free and baseline-agnostic, which enables it to generalize seamlessly to various generation tasks. Additionally, we provide an optional training scheme to explore further-improved possibilities. Experiments on both static multi-view generation and dynamic video generation show that CorrAdapter consistently improves spatiotemporal consistency and perceptual quality over strong baselines, offering a simple yet versatile drop-in approach to geometrically faithful multi-image diffusion. Code is available at https://github.com/SuhZhang/CorrAdapter.

*Corresponding author

1. Introduction

Generative diffusion models have demonstrated impressive capabilities in both multi-view synthesis [8, 11, 13, 24, 26, 38] and video generation [19, 25, 44, 49, 53]. These tasks require denoising multiple images jointly within each model timestep, where achieving consistency across generated frames or views remains a persistent challenge. Despite extensive training, such models still exhibit noticeable inconsistencies and visual artifacts, as illustrated in Figures 1 and 2(a).

To eliminate these inconsistencies, the generated details must be aligned across images according to their semantic and structural similarity. Intuitively, if the correspondences between different regions were accessible, they could serve as effective guidance to improve cross-image consistency. However, since all images are generated from pure Gaussian noise, no geometric or semantic priors such as depth maps or segmentation masks are available, making it difficult to identify explicit correspondences across images and then enforce the denoising process to align them.

How to get correspondences and align images before they are generated? This appears to be a classic chicken-and-egg problem. When no external prior knowledge is provided, a natural question arises: can useful information be mined directly from the intrinsic properties of the model itself. For instance, although multi-image diffusion models often exhibit inconsistencies, the generated images usually depict the same underlying scene. This observation leads to the following hypothesis: diffusion models have implicitly learned meaningful correspondences across images within intermediate noisy features.

Motivated by this hypothesis, we investigate whether multi-image diffusion models internally encode correspondences through their intrinsic noise features. Remarkably, we observe that, even before image synthesis begins, the intermediate feature space contains meaningful structural alignments across views. As illustrated in Figures 2(b) and 5, these *diffusion-native correspondences* emerge between semantically or geometrically similar regions, despite being embedded within noisy representations.

Building on this observation, we introduce **CorrAdapter**, a plug-and-play module that explicitly leverages these native correspondences during generation. CorrAdapter emphasizes feature alignment between matched regions across images throughout the denoising process, thereby enhancing spatiotemporal consistency, as shown in Figures 1 and 2(c). Specifically, CorrAdapter introduces a bypass branch for transformer blocks in the multi-image diffusion models. This branch encompasses a *native correspondence constructor* that builds correspondences from intermediate diffusion features using image feature matching approaches, and an *aligned area aggregator* that integrates features from only corresponding regions to prevent ambiguous information exchange. Through this design, CorrAdapter enforces consistency guided by diffusion-native correspondence, without requiring any external inputs or additional supervision. With its reliance solely on diffusion internal features, applicability across diverse generation settings, and lightweight plug-and-play structure, CorrAdapter serves as an efficient and versatile spatiotemporal consistency adapter, which is capable of aligning images even before they are fully generated. We sum up our contributions as follows:

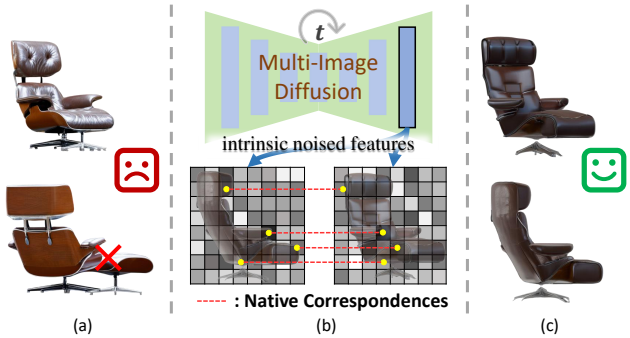

Figure 2. Consistency improvement of generated images with CorrAdapter. (a) Original multi-image generation results with inconsistent regions. (b) Native correspondences built by the diffusion model's intrinsic noise features before images are generated. (c) Aligned images after CorrAdapter is applied.

- We discover an innate property of multi-image diffusion models about building reliable correspondences with the intermediate-layer features, providing a matching prior for consistency constraints. These native correspondences help multi-image diffusion align images before they are generated.
- We propose a spatiotemporal consistency booster, CorrAdapter, for both static and dynamic multi-image generation. It builds correspondences with only the diffusion model itself at inference time, and emphasizes information aggregation on matching areas to ensure the spatiotemporal consistency of generated images.
- We design a simple yet effective structure for CorrAdapter to make it a training-free and plug-and-play adapter, fitting various tasks and methods. An optional training scheme is provided for further improvements.
- Experiments on both static multi-view and dynamic video generation show that CorrAdapter consistently improves spatiotemporal consistency over strong baselines.

2. Related Work

Spatiotemporal consistency is a fundamental challenge in multi-image generation, and many methods have been proposed to address it. Some previous works cast hope on the cross-image transformer to extract consistency prior from vast data, such as [2, 13, 20, 26, 38, 44, 53, 55]. This

approach is suitable for both static and dynamic scenes. But it requires large-scale, data-driven learning, and the implicit guidance is weak, still resulting in texture and structure drift when wide viewpoint changes or large temporal shifts. Then, researchers further seek the assistance of harder geometric constraints, including multi-view geometry [14, 24], epipolar constraint [16, 20, 30, 42, 54], and depth maps [10, 48], despite being only applicable to static scenes and requiring known images or depths as input. Whereas some video diffusions adopt optical flow estimation to model the movement of frames [15, 21, 32, 36] to fit in the dynamic scenes, and then correspondences are extracted with matching methods to constrain the consistency between aligned regions [3, 7, 22, 45, 47]. But still, they require video or key frames as input (such as video editing) and are not easily practical for pure generative tasks. In a nutshell, existing methods struggle badly with *a) applicability for both static and dynamic scenes*, and *b) eliminating reliance on extra inputs*, thereby they can not serve as the versatile consistency booster for various multi-image generation tasks. In this paper, we explore the native correspondences within the multi-image diffusion model itself, providing reliable guidance to improve consistency in a plug-and-play manner for a pure generation model on different tasks. Although the intrinsic correspondences in the diffusion model have been discovered in the single-image diffusion model [41], and its following works further improve the matching capabilities [28, 51]. But they still require a known image as input to reproduce intermediate features, while we explore the correspondences from a pure generative multi-image diffusion to align the generated images.

3. Method

The proposed CorrAdapter aims to align the images generated by multi-image diffusion models to enforce the spatiotemporal consistency. Given the idea introduced in Section 1 that guides the generative process with the diffusion-native correspondences, CorrAdapter boosts the consistency with two key steps: building correspondences from the diffusion model itself, and then modulating the information interaction across images according to these correspondences. Accordingly, as shown in Figure 3, CorrAdapter consists of two main components: **native correspondence constructor** and **aligned area aggregator**. These two simple modules can be integrated into various multi-image diffusion models, making CorrAdapter a plug-and-play adapter to improve consistency for both static and dynamic multi-image generation. We will detail CorrAdapter's structure in the following, together with an optional training scheme for further improvement if available.

3.1. Preliminaries

We first formulate the basics of multi-image diffusion models as the foundation for our method.

Multi-Image Diffusion Model The multi-image diffusion model discussed in this paper denotes a diffusion-based framework that generates multiple images of a certain scene within a single inference process. This includes static scenes, such as multi-view generation [26, 38], and dynamic scenes, such as video generation [25, 53]. Generally, these models aim to generate a set of N images $\boldsymbol{X} = \{\boldsymbol{x}^i | \boldsymbol{x}^i \in \mathbb{R}^{H_{\text{full}} \times W_{\text{full}} \times 3}\}_{i=1}^N$ based on a shared condition $\mathcal{C}$ (*e.g.*, a text prompt, a single reference image, *etc.*) that describes the scene. And the generation process is implemented by a latent diffusion model [13, 35] that predicts the latent representations $\boldsymbol{Z} = \{\boldsymbol{z}^i | \boldsymbol{z}^i \in \mathbb{R}^{H_{\text{latent}} \times W_{\text{latent}} \times C}\}_{i=1}^N$ encoded from the images $\boldsymbol{X}$ with VAE [18]. Specifically, the model learns a probabilistic distribution of all latent features $p_\theta(\boldsymbol{Z}_0|\mathcal{C}) := p_\theta(\boldsymbol{z}_0^1, ..., \boldsymbol{z}_0^N|\mathcal{C})$, which is typically formulated by a T-step Markov chain as the *reverse process*:

$$p_\theta(\boldsymbol{Z}_0|\mathcal{C}) = p(\boldsymbol{Z}_T) \prod_{t=1}^{T} p_\theta(\boldsymbol{Z}_{t-1}|\boldsymbol{Z}_t, \mathcal{C}), \tag{1}$$

$$p_\theta(\boldsymbol{Z}_{t-1}|\boldsymbol{Z}_t, \mathcal{C}) = \mathcal{N}(\boldsymbol{\mu}_\theta(\boldsymbol{Z}_t, t, \mathcal{C}), \sigma_t^2 \boldsymbol{I}), \tag{2}$$

where θ denotes the model parameters, $\boldsymbol{Z}_T \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ is a standard Gaussian noise. To learn $\boldsymbol{\mu}_\theta$, another Markov chain called *forward process* is constructed as:

$$q_\theta(\boldsymbol{Z}_t|\boldsymbol{Z}_{t-1}) = \mathcal{N}(\sqrt{1-\beta_t}\boldsymbol{Z}_{t-1}, \beta_t \boldsymbol{I}). \tag{3}$$

In Eqs. (2) and (3), σ_t, β_t are pre-defined variance and noise schedules. Furthermore, the mean prediction $\boldsymbol{\mu}_\theta(\boldsymbol{Z}_t, t, \mathcal{C})$ is usually transformed into a noise prediction:

$$\boldsymbol{\mu}_\theta(\boldsymbol{Z}_t, t, \mathcal{C}) = \frac{1}{\sqrt{1-\beta_t}} \left(\boldsymbol{Z}_t - \beta_t \boldsymbol{\epsilon}_\theta(\boldsymbol{Z}_t, t, \mathcal{C})\right). \tag{4}$$

The denoising network $\boldsymbol{\epsilon}_\theta$ is usually implemented by a powerful transformer-based architecture [33, 35, 43] which interacts the messages within the latent features $\boldsymbol{Z} = \{\boldsymbol{z}^i\}_{i=1}^N$ of all images, thereby generating results that are highly correlated. The noise $\boldsymbol{\epsilon}_\theta$ together with the model parameters θ are learned by minimizing the following loss:

$$\mathcal{L}_{\text{diffusion}} = \mathbb{E}_{t, \boldsymbol{Z}_0, \boldsymbol{\epsilon}}[\|\boldsymbol{\epsilon}_\theta(\boldsymbol{Z}_t, t, \mathcal{C}) - \boldsymbol{\epsilon}\|^2], \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}). \tag{5}$$

After elaborated training, $\boldsymbol{\epsilon}_\theta$ can be used to generate the latent features $\boldsymbol{Z}_0$ with Eqs. (1), (2) and (4). Then the images $\boldsymbol{X}_0$ can be recovered with a pre-trained VAE decoder [18]. In this simple pipeline without any additional inputs as geometric or semantic priors, we aim to leverage the diffusion-native correspondences to ensure spatiotemporal consistency among generated images before they are in fact produced.

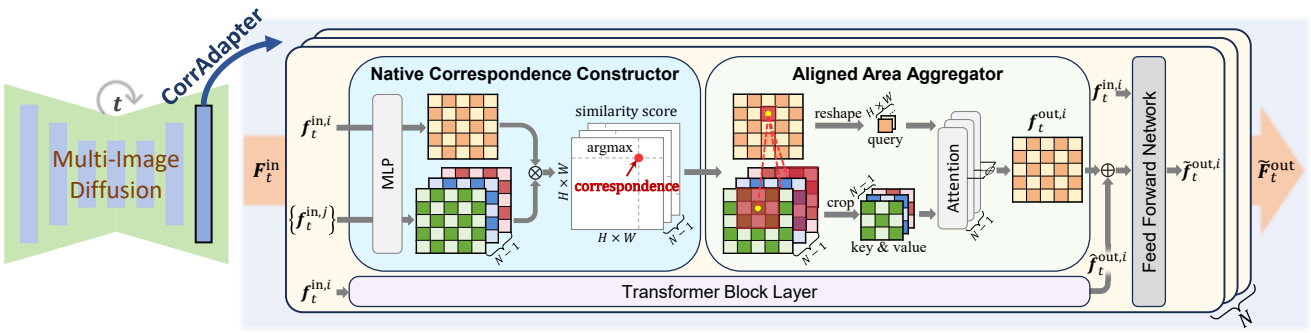

Figure 3. Pipeline of CorrAdapter. During each timestep in the multi-image diffusion model, CorrAdapter is juxtaposed with original transformer blocks at the highest resolution. It first builds native correspondences from the intrinsic features $\boldsymbol{F}_t^{\text{in}}$ for all image pairs, and then emphasizes information aggregation on aligned areas, updating the original output to $\tilde{\boldsymbol{F}}_t^{\text{out}}$.

3.2. Native Correspondence Constructor

The very first step of CorrAdapter is to build correspondences for all image pairs that will be generated. Given a pre-trained multi-image diffusion model, we propose a native correspondence constructor to find reliable correspondences without any additional inputs, and the structure is illustrated in Figure 3. Specifically, at timestep t, the input latent features $\boldsymbol{Z}_t$ are further characterized by a set of transformer blocks in the denoising network. Take a certain block as an example, its input is some condensed representations $\boldsymbol{F}_t^{\text{in}} = \{\boldsymbol{f}_t^{\text{in},i} | \boldsymbol{f}_t^{\text{in},i} \in \mathbb{R}^{H\times W\times D}\}_{i=1}^N$ of $\boldsymbol{Z}_t$. The transformer then updates these representations to $\boldsymbol{F}_t^{\text{out}}$ with Attention Mechanism [43] to fully consider the information in each image, which is formulated as:

$$\hat{\boldsymbol{F}}_t^{\text{out}} = \mathcal{A}(\boldsymbol{F}_t^{\text{in}}, \boldsymbol{F}_t^{\text{in}}) = \text{Softmax}\left(\frac{\boldsymbol{Q}_t \boldsymbol{K}_t{}^T}{\sqrt{D}}\right)\boldsymbol{V}_t, \quad (6)$$

$$\boldsymbol{F}_t^{\text{out}} = \boldsymbol{F}_t^{\text{in}} + \text{FFN}(\boldsymbol{F}_t^{\text{in}} \| \hat{\boldsymbol{F}}_t^{\text{out}}), \quad (7)$$

where $\boldsymbol{Q}_t = \boldsymbol{W}_Q \boldsymbol{F}_t^{\text{in}}$, $\boldsymbol{K}_t = \boldsymbol{W}_K \boldsymbol{F}_t^{\text{in}}$, $\boldsymbol{V}_t = \boldsymbol{W}_V \boldsymbol{F}_t^{\text{in}}$ are the query, key, and value matrices, and $\boldsymbol{Q} \in \mathbb{R}^{(N\times H\times W)\times D}$, $\boldsymbol{K} \in \mathbb{R}^{(N\times H\times W)\times D}$, $\boldsymbol{V} \in \mathbb{R}^{(N\times H\times W)\times D}$, respectively. FFN$(\cdot)$ is the feed-forward network, and $\|$ denotes the concatenation operation. Considering the Softmax$(\cdot)$ operation, the attention weights reflect the similarity between intermediate features $\boldsymbol{F}_t^{\text{in}} = \{\boldsymbol{f}_t^{\text{in},i}\}_{i=1}^N$ of each image. Thus, these weights can be regarded as the similarity scores and naturally used to depict the correspondence between different images. For instance, the similarity socre $\boldsymbol{s}_t^{i,j} \in \mathbb{R}^{(H\times W)\times(H\times W)}$ for image pair $(i,j), i \neq j$ at timestep t is calculated as:

$$\boldsymbol{s}_t^{i,j} = \text{Softmax}\left(\frac{\boldsymbol{q}_t^i \boldsymbol{k}_t^{j\,T}}{\sqrt{D}}\right), \quad (8)$$

where $\boldsymbol{q}_t^i = \boldsymbol{Q}_t[i], \boldsymbol{q}_t^i \in \mathbb{R}^{(H\times W)\times D}$ and $\boldsymbol{k}_t^j = \boldsymbol{K}_t[j], \boldsymbol{k}_t^j \in \mathbb{R}^{(H\times W)\times D}$ are the i-th and j-th slices of $\boldsymbol{Q}_t$ and $\boldsymbol{K}_t$, respectively. Then traditional image matching methods [31] can be applied on the similarity scores to build correspondences, such as the nearest neighbor matching:

$$\boldsymbol{c}_t^{i,j} = \text{argmax}_{l\in\{1,\dots,H\times W\}} \boldsymbol{s}_t^{i,j}[:,l], \quad (9)$$

where $[:,l]$ means the l-th column of the matrix, $\boldsymbol{c}_t^{i,j} \in \mathbb{R}^{H\times W}$ indicates the nearest neighbor index of each i-th features $\boldsymbol{f}_t^{\text{in},i}$ in the j-th features $\boldsymbol{f}_t^{\text{in},j}$, reflecting the spatial correspondence of image pair (i,j). Thus, all correspondences $\boldsymbol{C}_t \in \mathbb{R}^{N\times(N-1)\times H\times W}$ can be obtained by stacking $\boldsymbol{c}_t^{i,j}$ for all image pairs (i,j) at timestep t. By re-using the attention weights in Eq. (6), we can easily build the correspondences with Eqs. (8) and (9) while avoiding redundant computations.

Following the above process, we can build the native correspondences with the intrinsic intermediate features in the diffusion model at every timestep. Then these correspondences can indicate the matching areas between different images so that we can modulate the information interaction across images accordingly.

3.3. Aligned Area Aggregator

The correspondences built in the previous section identify the most geometrically or semantically related region between the features of different images. To enhance spatiotemporal consistency, a natural strategy is to strengthen information exchange between these corresponding features within the transformer blocks. To this end, we propose an aligned area aggregator as shown in Figure 3 that modulates feature integration by emphasizing matching regions while suppressing others, thereby reducing ambiguity. Once given the correspondences $\boldsymbol{c}_t^{i,j}$ of image pair (i,j), for each feature $\boldsymbol{f}_t^{\text{in},i}[k] \in \mathbb{R}^D, k = 1,\dots,H\times W$, we crop a $(2r+1)\times(2r+1)$ window with radius r on features $\boldsymbol{f}_t^{\text{in},j}$ accroding to $\boldsymbol{c}_t^{i,j}[k] \in \mathbb{R}$ to form the aggregated

feature $\hat{\boldsymbol{f}}_t^i[k]$:

$$\hat{\boldsymbol{f}}_t^i[k] = \text{Softmax}\left(\frac{\boldsymbol{w}_q \boldsymbol{f}_t^{\text{in},i}[k](\boldsymbol{w}_k \boldsymbol{f}_{t,\text{crop-}k}^{\text{in},j})^T}{\sqrt{D}}\right)\boldsymbol{w}_v \boldsymbol{f}_{t,\text{crop-}k}^{\text{in},j}, \tag{10}$$

where $\boldsymbol{f}_{t,\text{crop-}k}^{\text{in},j} = \boldsymbol{f}_t^{\text{in},j}\left[\boldsymbol{c}_t^{i,j}[k]-r : \boldsymbol{c}_t^{i,j}[k]+r, \boldsymbol{c}_t^{i,j}[k]-r : \boldsymbol{c}_t^{i,j}[k]+r\right]$, $\boldsymbol{f}_{t,\text{crop-}k}^{\text{in},j} \in \mathbb{R}^{(2r+1)^2 \times D}$ is the cropped corresponding area for the k-th feature of $\boldsymbol{f}_t^{\text{in},j}$. Comparing Eq. (10) with Eq. (6), we observe that the aggregated feature $\hat{\boldsymbol{f}}_t^i[k]$ is a weighted sum of the features within the cropped corresponding area, where the weights are precisely the attention weights computed from the query and key values of the cropped features. Then Eq. (10) can be rewritten as:

$$\hat{\boldsymbol{f}}_t^i[k] = \text{Softmax}\left(\frac{\boldsymbol{q}_t^i[k]\boldsymbol{k}_{t,\text{crop-}k}^{j\,T}}{\sqrt{D}}\right)\boldsymbol{v}_{t,\text{crop-}k}^j, \tag{11}$$

where $\boldsymbol{k}_{t,\text{crop-}k}^j = \boldsymbol{k}_t^j\left[\boldsymbol{c}_t^{i,j}[k]-r : \boldsymbol{c}_t^{i,j}[k]+r, \boldsymbol{c}_t^{i,j}[k]-r : \boldsymbol{c}_t^{i,j}[k]+r\right]$ and the same for $\boldsymbol{v}_{t,\text{crop-}k}^j$. In this way, we can again reuse the attention weights in Eq. (6) to calculate the aggregated features, thereby avoiding heavy computational cost.

By repeatedly applying Eqs. (11) and (13) to all features in $\boldsymbol{F}_t^{\text{in}}$, we obtain the aggregated features $\hat{\boldsymbol{F}}_t$, where information is integrated only within mutually consistent regions. To further enhance feature consistency inside the transformer blocks, we add the aggregated features $\hat{\boldsymbol{F}}_t$ to the original output $\hat{\boldsymbol{F}}_t^{\text{out}}$ in Eq. (6):

$$\check{\boldsymbol{F}}_t^{\text{out}} = \eta\hat{\boldsymbol{F}}_t^{\text{out}} + (1-\eta)\hat{\boldsymbol{F}}_t, \tag{12}$$

where η is a hyper-parameter to balance the importance of the aligned-area aggregated features and the original ones. Then, similar to the attention mechanism in Eq. (7), we apply a feed-forward network to the aggregated features:

$$\tilde{\boldsymbol{F}}_t^{\text{out}} = \check{\boldsymbol{F}}_t^{\text{out}} + \text{FFN}(\boldsymbol{F}_t^{\text{in}} \| \check{\boldsymbol{F}}_t^{\text{out}}), \tag{13}$$

which will serve as an updated input to the next block. With Eqs. (11), (12), and (13), the aligned area aggregator enforces the diffusion model to focus on the matching regions and suppress the non-matching ones, thereby enhancing the spatiotemporal consistency of the generated images.

Here, we have provided a detailed outline of the component structures corresponding to the two steps of correspondence construction and consistency enhancement within the CorrAdapter. To be regarded as a plug-and-play adapter for various multi-image diffusion backbones, we will then introduce the inference details for effectively utilizing CorrAdapter in a variety of multi-image generation tasks.

3.4. Inference Details

As shown in Figure 3, CorrAdapter is just injected into the diffusion model as a bypass branch of the existing transformer blocks, and its output is added to the original one as seen in Eq. (12). However, the advanced multi-image diffusion models often adopt cascaded or parallel transformers to enhance its performance [13, 38]. We recommend to integrate CorrAdapter into the transformer blocks that model the multi-image interaction so that it can reuse the attention weights in Eq. (6) to calculate both the native correspondences and the aggregated features. Otherwise, juxtaposing CorrAdapter with the vanilla transformer inherited from single-image diffusion models as seen in many inchoate works [24] is available, albeit with extra computational cost for corss-image similarity calculations. A practicable trick is to share the native correspondences between several adjacent timesteps to reduce the additional computations. We will detail this advice in Section 4.3. Additionally, since the native correspondences are built from the intermediate features in the latent space, we consider only the highest level of the transformer to make correspondences more fine-grained. All learnable parameters of CorrAdapter are initialized with the same values as the existing transformer blocks it is injected into, and are then kept frozen during inference. Such a training-free strategy makes CorrAdapter easily implementable and compatible with various multi-image diffusion models.

3.5. Training Scheme

To enable CorrAdapter to be rapidly and widely adopted across different multi-image generation models and tasks, we design it as a training-free paradigm. Moreover, for scenarios that demand higher performance, we also provide an optional training scheme tailored to specific use cases. Specifically, we fine-tune the original diffusion model to learn more consistency intermediate features $\boldsymbol{F}_t^{\text{in}} = \{\boldsymbol{f}_t^{\text{in},i}\}_{i=1}^N$, with the hope to establish more reliable native correspondences in Eq. (8). This is achieved by adding a consistency loss:

$$\mathcal{L}_{\text{consistency}} = \sum_{(i,j),i\neq j} \|\boldsymbol{f}_t^{\text{in},i}[k] - \boldsymbol{f}_t^{\text{in},j}\left[\dot{\boldsymbol{c}}^{i,j}[k]\right]\|^2, \tag{14}$$

where $\dot{\boldsymbol{c}}^{i,j} \in \mathbb{R}^{H\times W}$ indicates the referenced correspondences captured from ground truth image pair (i,j). These correspondences can be constructed with off-the-shelf image matching algorithms [31, 56]. We adopt LoRA-based fine-tuning [9] on all transformer blocks to accelerate training, which introduces low-rank adaptation weights into Eqs. (6) and (7). Loss function (14) is also applied in each training step as the same as Eq. (5), so the total loss is:

$$\mathcal{L} = \mathcal{L}_{\text{diffusion}} + \lambda\mathcal{L}_{\text{consistency}}, \tag{15}$$

where λ is the hyper-parameter to balance loss functions.

Note that the training process relies on explicit consistency constraints between diffusion features and ground-truth images. This is reasonable for certain tasks such as

Table 1. Results on image-conditioned multi-view generation.

Method	Single-Image Quality			Geometric Consistency					
	PSNR↑	SSIM↑	LPIPS↓	cPSNR↑	cSSIM↑	cLPIPS↓	CD↓	depth↓	MEt3R↓
SyncDreamer [24]	19.24	0.8215	0.1557	26.28	0.8817	0.0634	**2.66**	14.15	0.1656
SyncDreamer+CorrAdapter	**19.72**	**0.8310**	**0.1472**	**27.20**	**0.8926**	**0.0572**	2.67	**12.60**	**0.1529**
MVAdapter [24]	23.15	0.8761	0.1276	18.75	0.7201	0.2519	9.75	73.47	0.2116
MVAdapter+CorrAdapter	23.82	0.8829	0.1235	19.68	0.7396	0.2371	**9.20**	68.62	0.2036
MVAdapter+CorrAdapter*	**24.05**	**0.8866**	**0.1220**	**20.47**	**0.7634**	**0.2256**	9.33	**67.47**	**0.1955**

Table 2. Results on text-conditioned multi-view generation.

Method	Single-Image Quality			Geometric Consistency					
	FID↓	IS↑	CLIP-Score↑	cPSNR↑	cSSIM↑	cLPIPS↓	CD↓	depth↓	MEt3R↓
MVAdapter [13]	24.20	15.22	33.10	14.09	0.5797	0.3513	12.70	79.84	0.3017
MVAdapter+CorrAdapter	**23.42**	15.96	33.17	14.29	0.5786	0.3418	12.63	79.72	0.2986
MVAdapter+CorrAdapter*	24.17	**17.07**	**33.52**	**15.27**	**0.6275**	**0.3085**	**11.61**	**77.94**	**0.2701**

image-conditioned multi-view generation, but inappropriate for tasks whose outputs are inherently diverse. We therefore propose transplanting the LoRA modules from models for which this training scheme is applicable to those for which it is not, thereby inheriting their consistency-enhancing capability. However, this transfer is only feasible when the network architectures are identical, and we will give an example in Section 4.1. Consequently, the training scheme described in this section is optional, whereas the training-free strategy can be applied broadly across different models.

4. Experiments

In this section, we evaluate the performance of CorrAdapter on various multi-image generation tasks with several strong baselines to prove CorrAdapter's effectiveness in improving the spatiotemporal consistency of generated images, and also to verify its generalization ability to various scenes. Concretely, we first conduct experiments on static multi-view generation, and then on dynamic video generation. And we also provide a detailed analysis to further understand the working mechanism of CorrAdapter.

4.1. Static Scene: Multi-View Generation

We choose multi-view generation as the static scene generation task to evaluate the performance of CorrAdapter, including both the image-conditioned multi-view generation ($\mathcal{C}$ in Eq. (1) is a single reference image) and text-conditioned one ($\mathcal{C}$ is a text prompt).

Evaluation protocols. For image-conditioned multi-view generation, common metrics like `PSNR`, `SSIM`, and `LPIPS` are adopted to evaluate the quality of every single image [23, 24]. Although these metrics somehow reflect the coherence of different generated views, they can not intuitively assess the geometric consistency. Therefore, we adopt the proposed 3D consistency metrics in MVGBench [50]. MVGBench splits the synthesized imaged into two parts, and reconstructs 3D models from these parts respectively with 3DGS [17]. The metrics include `cPSNR`, `cSSIM`, and `cLPIPS`, which are calculated by two images with the same camera parameters but rendered from different reconstructed 3D models, and Chamfer Distance (`CD`) together with rendered depth error (`depth`) that evaluate the consistency of the 3D models. Additionally, we adopt `MEt3R` [1], which uses a more powerful construction method DUSt3R [46] to model the 3D scene, as an auxiliary indicator. Similar to [24], our evaluation is conducted on the Google Scanned Objects (GSO) dataset [6] with 100 randomly selected scenes. For text-conditioned multi-view generation, we adopt typical single-image metrics like `FID`, `IS`, and `CLIP-Score` [13, 52], and also choose the same 3D consistency metrics with image-conditioned methods. The evaluation is performed on the Objaverse dataset [4] with 1000 prompts randomly selected referring to [13].

Comparison with baselines. We choose SyncDreamer [24] and MVAdapter [13] as the image-conditioned baselines for comparison, and the text-to-multi-view variant of MVAdapter for text-conditioned generation. SyncDreamer synthesizes 16 views with 30° elevation, while MVAdapter generates 6 views with 0° elevation. Results are reported in Tables 1 and 2. We additionally mark methods that use the training scheme with superscript *. CorrAdapter improves different baselines, and is suitable for both image-conditioned and text-conditioned generation. The geometric consistency of generated images is significantly boosted, which in turn assists the single-image quality. And the proposed training scheme further enhances the performance with both condi-

Figure 4. Visualizations of CorrAdapter on different multi-image generation tasks. CorrAdapter fixes different kinds of inconsistencies of generated images, including distorted structure, chaos, content lost, and so on.

tions while the models share the same architecture. Apart from 1, we provide further visualizations in Figure 4 and in the Appendix to support the superiority of CorrAdapter.

4.2. Dynamic Scene: Video Generation

To assess CorrAdapter on dynamic scene generation, we select the video generation task that tries to produce a sequence of images with a text prompt as the condition.

Evaluation protocols. Following many previous video generation tasks, we adopt VBench [12] to evaluate the performance. It provides standard metrics on numerous dimensions, and we evaluate the quality of generated images (frames) with a total of 10 dimensions, covering temporal quality, single-frame quality, and text-video consistency.

Comparison with baselines. For video generation, we choose Wan2.1-1.3B [44] as the baseline (abbreviated as Wan2.1). Results are reported in Table 3. With the simple training-free CorrAdapter, the consistency of subjects and backgrounds improves excellently, proving its effectiveness on spatiotemporally consistent video generation. This also helps to produce better coherence with the text prompt. And due to the concerns about consistency, the dynamic degree decreases foreseeably, but it does not affect the improvement of the overall video quality. We also provide more visualizations, as shown in Figure 4 and in the Appendix.

4.3. Implementation Details

To make CorrAdapter a versatile adapter and applicable for different methods, we use a matching threshold of 0.05 instead of the poor nearest neighbor method in Eq. (9), and set $r = 3$ in Eq. (11). η in Eq. (12) is set as $\eta = 0.1$ by default, while for text-conditioned multi-view generation, it is increased to 0.8. And to preserve the diversity of text-conditioned generation, we apply CorrAdapter only to the first 10 timesteps for Table 2 and the first 15 timesteps for Table 3, while for image-conditioned methods it is applied at all timesteps. Additionally, for SyncDreamer, due to its lack of the multi-view transformer, we cannot reuse the attention weights in Eqs. (8) and (11), yielding a large computational burden to calculate the Softmax operation. Thus, we keep the native correspondences and attention weights, updating them every 5 timesteps, which provides a practicable approach to apply CorrAdapter on such a backbone. And for MVAdapter that uses row-wise attention to model the multi-view transformer, we crop local features in Eq. (11) only on the corresponding rows to avoid redundant calculations. All inferences are conducted on a single RTX 6000 GPU with 48 GB VRAM. For the training

Table 3. Results on text-conditioned video generation.

Method	Temporal Quality					Single-Frame Quality		Text-Video Consistency		
	Subject Consistency	Background Consistency	Temporal Flickering	Motion Smoothness	Dynamic Degree	Aesthetic Quality	Imaging Quality	Scene	Appearance Style	Overall Consistency
Wan2.1 [44]	0.9536	0.9626	**0.9940**	0.9851	**0.5556**	0.6016	**0.6660**	0.2202	0.2017	0.2275
Wan2.1+CorrAdapter	**0.9715**	**0.9696**	0.9931	**0.9857**	0.5139	**0.6071**	0.6560	**0.2878**	**0.2047**	**0.2320**

scheme, we adopt an efficient and reliable matching algorithm, LoFTR [40], to establish supervision for Eq. (14). And λ is set as 0.1 in Eq. (15). Different from typical fine-tuning, we first fine-tune the original model with LoRA [9] applying only the new loss function to learn more consistency intermediate features, then add our CorrAdapter structure on the fine-tuned model. We will discuss this strategy in Section 4.4. The fine-tuning process is conducted with the same settings as [13] for 1 epoch, which takes about 1 day on 4 RTX 6000 GPUs with the help of DeepSpeed [34]. Find more details in the Appendix.

4.4. Analysis

We further analyze the native correspondences and the optional training scheme of CorrAdapter in this section, together with the computational consumption statistics.

Are native correspondences reliable? To prove our claim about the existence of reliable native correspondences in pre-trained multi-image diffusion models, we use these correspondences to match the generated images. We also add some practical matching baselines like SIFT [27] and SuperPoint [5] for comparison. Several results are drawn in Figure 5. With the epipolar distance threshold set as 5 pixels, we mark the num of correct matches, total matches, and the matching accuracy in the image pairs. The native correspondences rival the matching baselines in both correspondence number and matching accuracy, providing CorrAdapter with strong reliability to guide the information interaction and align the generated images. Find more results in the Appendix.

An alternative training scheme? We have tried to train CorrAdapter directly by fine-tuning the original model with both the consistency loss and the two components altogether. The performance is bad, getting only 23.65, 0.8812, 0.1248 on PSNR, SSIM, and LPIPS for image-conditioned multi-view generation, even worse than the training-free CorrAdapter. We believe that the process of finding matches and cropping may clutter gradient backpropagation, jeopardizing training convergence. Therefore, the proposed training strategy that optimizes feature consistency before adding the matching module is more reasonable.

Does CorrAdapter cost a lot? We measure the inference time (`Time`), computational quantity (`Flops`), number of parameters (`Param`), and peak memory usage (`Mem`) for the baseline and the model with CorrAdapter. Results in Table 4 show that CorrAdapter enhances the consistency of generated images without substantial computational overhead.

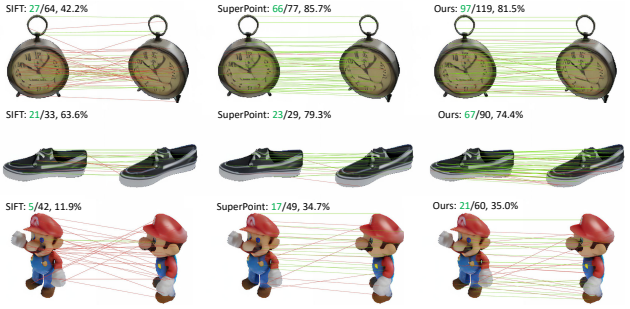

Figure 5. Visualizations of native correspondences and several matching baselines on generated images. The green lines indicate correct correspondences, and the red lines indicate incorrect ones.

Table 4. Runtime and resource usage.

Method	Time(s)	Flops(P)	Param(G)	Mem(GB)
MVAdapter [13]	33.42	2.71	4.29	15.27
MVAdapter+CorrAdapter	39.83	2.73	4.30	20.86
MVAdapter+CorrAdapter*	40.60	2.74	4.31	20.88

5. Conclusion

This paper presents CorrAdapter, a plug-and-play spatiotemporal consistency adapter for multi-image diffusion models to align images before they are generated. It discovers and exploits the *diffusion-native* correspondences built by the diffusion model's intrinsic intermediate features, providing a reliable matching prior for more effective information interaction and aggregation within the aligned areas. Due to the characteristics of not being dependent on any auxiliary inputs and hard geometric constraints, CorrAdapter serves as a versatile adapter for both static and dynamic multi-image generation tasks, and can be easily integrated into various baselines in a training-free manner. And an optional training scheme is also provided for some specific models. Overall, CorrAdapter explores a consistency constraint applicable to diverse multi-image diffusion architectures, which is expected to enhance a variety of downstream tasks utilizing this paradigm.

Acknowledgement

This project is supported by the Ministry of Education, Singapore, under its Academic Research Fund Tier 2 (Award Number: MOE-T2EP20122-0006), and National Research Foundation, Singapore, and Cyber Security Agency of Singapore under its National Cybersecurity R&D Programme and CyberSG R&D Cyber Research Programme Office (Award: CRPO-GC1-NTU-002).

References

[1] Mohammad Asim, Christopher Wewer, Thomas Wimmer, Bernt Schiele, and Jan Eric Lenssen. Met3r: Measuring multi-view consistency in generated images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6034–6044, 2025. 6

[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 Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023. 2

[3] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 22560–22570, 2023. 3

[4] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023. 6, 12

[5] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 224–236, 2018. 8

[6] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *Proceedings of the International Conference on Robotics and Automation*, pages 2553–2560, 2022. 6, 12

[7] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. In *Proceedings of the International Conference on Learning Representations*, pages 1–13, 2024. 3

[8] Lukas Höllein, Aljaž Božič, Norman Müller, David Novotny, Hung-Yu Tseng, Christian Richardt, Michael Zollhöfer, and Matthias Nießner. Viewdiff: 3d-consistent image generation with text-to-image models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5043–5052, 2024. 2

[9] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations*, pages 1–13, 2022. 5, 8, 12

[10] Hanzhe Hu, Zhizhuo Zhou, Varun Jampani, and Shubham Tulsiani. Mvd-fusion: Single-view 3d via depth-consistent multi-view generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9698–9707, 2024. 3

[11] Yingcheng Hu, Haowen Gong, Chuanguang Yang, Zhulin An, Yongjun Xu, and Songhua Liu. Multianimate: Pose-guided image animation made extensible. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–10, 2026. 2

[12] 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 Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 7, 12

[13] Zehuan Huang, Yuan-Chen Guo, Haoran Wang, Ran Yi, Lizhuang Ma, Yan-Pei Cao, and Lu Sheng. Mv-adapter: Multi-view consistent image generation made easy. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 16377–16387, 2025. 2, 3, 5, 6, 8, 12

[14] Yoonwoo Jeong, Jinwoo Lee, Chiheon Kim, Minsu Cho, and Doyup Lee. Nvs-adapter: Plug-and-play novel view synthesis from a single image. In *Proceedings of the European Conference on Computer Vision*, pages 449–466, 2024. 3

[15] Wonjoon Jin, Qi Dai, Chong Luo, Seung-Hwan Baek, and Sunghyun Cho. Flovd: Optical flow meets video diffusion model for enhanced camera-controlled video synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2040–2049, 2025. 3

[16] Yash Kant, Aliaksandr Siarohin, Ziyi Wu, Michael Vasilkovsky, Guocheng Qian, Jian Ren, Riza Alp Guler, Bernard Ghanem, Sergey Tulyakov, and Igor Gilitschenski. Spad: Spatially aware multi-view diffusers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10026–10038, 2024. 3

[17] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4):139, 2023. 6

[18] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *Proceedings of the International Conference on Learning Representations*, pages 1–14, 2014. 3

[19] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. 2, 12, 13

[20] Peng Li, Yuan Liu, Xiaoxiao Long, Feihu Zhang, Cheng Lin, Mengfei Li, Xingqun Qi, Shanghang Zhang, Wei Xue, Wenhan Luo, et al. Era3d: High-resolution multiview diffusion using efficient row-wise attention. *Advances in Neural Information Processing Systems*, 37:55975–56000, 2024. 2, 3

[21] Feng Liang, Bichen Wu, Jialiang Wang, Licheng Yu, Kunpeng Li, Yinan Zhao, Ishan Misra, Jia-Bin Huang, Peizhao

Zhang, Peter Vajda, et al. Flowvid: Taming imperfect optical flows for consistent video-to-video synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8207–8216, 2024. 3

[22] Jingyun Liang, Yuchen Fan, Kai Zhang, Radu Timofte, Luc Van Gool, and Rakesh Ranjan. Movideo: Motion-aware video generation with diffusion model. In *Proceedings of the European Conference on Computer Vision*, pages 56–74, 2024. 3

[23] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9298–9309, 2023. 6

[24] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. In *Proceedings of the International Conference on Learning Representations*, pages 1–22, 2024. 2, 3, 5, 6, 12

[25] Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, et al. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402.17177*, 2024. 2, 3

[26] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9970–9980, 2024. 2, 3

[27] David G Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. 8

[28] Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. *Advances in Neural Information Processing Systems*, 36: 47500–47510, 2023. 3

[29] Tiange Luo, Chris Rockwell, Honglak Lee, and Justin Johnson. Scalable 3d captioning with pretrained models. *Advances in Neural Information Processing Systems*, 36: 75307–75337, 2023. 12

[30] Yihang Luo, Shangchen Zhou, Yushi Lan, Xingang Pan, and Chen Change Loy. 3denhancer: Consistent multi-view diffusion for 3d enhancement. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 16430–16440, 2025. 3

[31] Jiayi Ma, Xingyu Jiang, Aoxiang Fan, Junjun Jiang, and Junchi Yan. Image matching from handcrafted to deep features: A survey. *International Journal of Computer Vision*, 129(1):23–79, 2021. 4, 5

[32] Hyelin Nam, Jaemin Kim, Dohun Lee, and Jong Chul Ye. Optical-flow guided prompt optimization for coherent video generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7837–7846, 2025. 3

[33] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *Proceedings of the International Conference on Learning Representations*, pages 1–13, 2024. 3

[34] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining*, pages 3505–3506, 2020. 8, 12

[35] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 3

[36] Fengyuan Shi, Jiaxi Gu, Hang Xu, Songcen Xu, Wei Zhang, and Limin Wang. Bivdiff: A training-free framework for general-purpose video synthesis via bridging image and video diffusion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7393–7402, 2024. 3

[37] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023. 12, 13

[38] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. In *Proceedings of the International Conference on Learning Representations*, pages 1–22, 2024. 2, 3, 5

[39] Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. In *Proceedings of the International Conference on Learning Representations*, pages 1–22, 2024. 12, 13

[40] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8922–8931, 2021. 8, 12

[41] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *Advances in Neural Information Processing Systems*, 36:1363–1389, 2023. 3

[42] Hung-Yu Tseng, Qinbo Li, Changil Kim, Suhib Alsisan, Jia-Bin Huang, and Johannes Kopf. Consistent view synthesis with pose-guided diffusion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 16773–16783, 2023. 3

[43] 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*, 30:1–11, 2017. 3, 4

[44] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 2, 7, 8, 12

[45] Jiangshan Wang, Yue Ma, Jiayi Guo, Yicheng Xiao, Gao Huang, and Xiu Li. Cove: Unleashing the diffusion feature correspondence for consistent video editing. *Advances in Neural Information Processing Systems*, 37:96541–96565, 2024. 3

[46] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024. 6

[47] Shengzhi Wang, Yingkang Zhong, Jiangchuan Mu, Kai Wu, Mingliang Xiong, Wen Fang, Mingqing Liu, Hao Deng, Bin He, Gang Li, et al. Align-a-video: Deterministic reward tuning of image diffusion models for consistent video editing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2074–2083, 2025. 3

[48] Zhen Wang, Qiangeng Xu, Feitong Tan, Menglei Chai, Shichen Liu, Rohit Pandey, Sean Fanello, Achuta Kadambi, and Yinda Zhang. Mvdd: Multi-view depth diffusion models. In *Proceedings of the European Conference on Computer Vision*, pages 236–253, 2024. 3

[49] Yunfeng Wu, Jiayi Song, Zhenxiong Tan, Zihao He, and Songhua Liu. Freeswim: Revisiting sliding-window attention mechanisms for training-free ultra-high-resolution video generation. *arXiv preprint arXiv:2511.14712*, 2025. 2

[50] Xianghui Xie, Chuhang Zou, Meher Gitika Karumuri, Jan Eric Lenssen, and Gerard Pons-Moll. Mvgbench: Comprehensive benchmark for multi-view generation models. pages 8207–8218, 2025. 6

[51] Fei Xue, Sven Elflein, Laura Leal-Taixé, and Qunjie Zhou. Matcha: Towards matching anything. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 27081–27091, 2025. 3

[52] Yuanbo Yang, Jiahao Shao, Xinyang Li, Yujun Shen, Andreas Geiger, and Yiyi Liao. Prometheus: 3d-aware latent diffusion models for feed-forward text-to-3d scene generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2857–2869, 2025. 6

[53] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *Proceedings of the International Conference on Learning Representations*, pages 1–30, 2024. 2, 3

[54] Botao Ye, Sifei Liu, Xueting Li, Marc Pollefeys, and Ming-Hsuan Yang. Synthesizing consistent novel views via 3d epipolar attention without re-training. In *Proceedings of the International Conference on 3D Vision*, pages 337–346, 2025. 3

[55] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–18, 2025. 2

[56] Shihua Zhang, Zizhuo Li, Kaining Zhang, Yifan Lu, Yuxin Deng, Linfeng Tang, Xingyu Jiang, and Jiayi Ma. Deep learning reforms image matching: A survey and outlook. *arXiv preprint arXiv:2506.04619*, 2025. 5