

Panorama Generation From NFoV Image Done Right

Dian Zheng^{1,*} Cheng Zhang² Xiao-Ming Wu¹
Cao Li^{3,✉} Chengfei Lv³ Jian-Fang Hu¹ Wei-Shi Zheng^{1,4,✉}

¹Sun Yat-sen University ²Monash University ³Alibaba Group

⁴Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China

<https://isee-laboratory.github.io/PanoDecouple/>

Abstract

Generating 360-degree panoramas from narrow field of view (NFoV) image is a promising computer vision task for Virtual Reality (VR) applications. Existing methods mostly assess the generated panoramas with InceptionNet or CLIP based metrics, which tend to perceive the image quality and is **not suitable for evaluating the distortion**. In this work, we first propose a distortion-specific CLIP, named *Distort-CLIP* to accurately evaluate the panorama distortion and discover the “**visual cheating**” phenomenon in previous works (i.e., tending to improve the visual results by sacrificing distortion accuracy). This phenomenon arises because prior methods employ a single network to learn the distinct panorama distortion and content completion at once, which leads the model to prioritize optimizing the latter. To address the phenomenon, we propose **PanoDecouple**, a decoupled diffusion model framework, which decouples the panorama generation into distortion guidance and content completion, aiming to generate panoramas with both accurate distortion and visual appeal. Specifically, we design a *DistortNet* for distortion guidance by imposing panorama-specific distortion prior and a modified condition registration mechanism; and a *ContentNet* for content completion by imposing perspective image information. Additionally, a distortion correction loss function with *Distort-CLIP* is introduced to constrain the distortion explicitly. The extensive experiments validate that *PanoDecouple* surpasses existing methods both in distortion and visual metrics.

1. Introduction

360-degree panorama generation from narrow field of view (NFoV) image aims to outpaint the partial panorama while

CVPR2022
OmniDreamer
Distort-FID: 0.52
FID: 75.14

ACMMM2023
PanoDiff
Distort-FID: 2.68
FID: 63.49

ECCV2024
2S-ODIS
Distort-FID: 8.42
FID: 72.23

PanoDecouple (Ours)
Distort-FID: 0.92
FID: 62.19

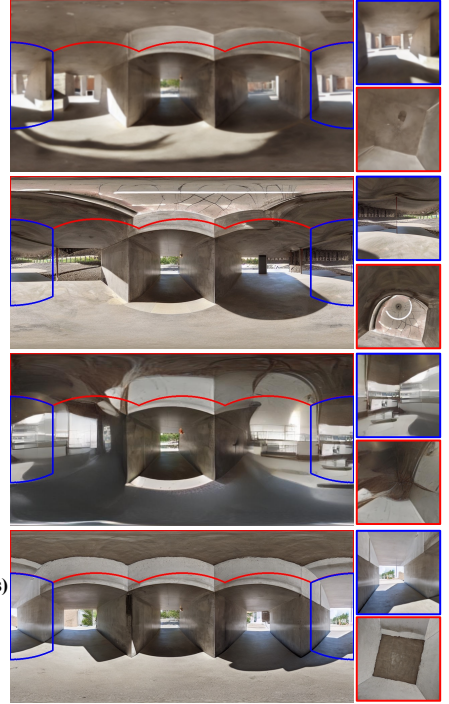


Figure 1. The image quality and distortion accuracy of existing methods and ours by FID and Distort-FID (ours) respectively. We project two regions in panorama (signed in corresponding color) into perspective image to show the distortion accuracy of existing methods (i.e., no distortion and natural layout in perspective image means good results). Recent methods improve the image quality while significantly ruining the distortion. We named it “visual cheating” phenomenon. Zoom in for best view.

maintaining accurate distortion and ensuring consistency in content, style with the NFoV image simultaneously, which has a wide range of applications in Virtual Reality (VR) and 3D scene generation [2, 57].

Existing methods mainly utilize InceptionNet [38] or CLIP [34] based evaluation metrics (e.g., FID, IS, etc.) to

* The work is done when Dian Zheng is an intern at Alibaba Group.

✉ corresponding authors. [Code](#) is available

validate the generation performance. However, these models tend to perceive the image quality while ignoring the distortion as the feature similarity between panoramas with different contents is lower than panorama and perspective image with the same content (The detail is in Sec 3). To address this, we propose a distortion-perception CLIP, named Distort-CLIP, which is tuned by different distortion types of data in contrastive learning mechanism. We substitute the InceptionNet used in FID with our Distort-CLIP and compute the Distort-FID of existing methods, observing the “visual cheating” phenomenon as shown in Fig 1. Surprisingly, OmniDreamer [1] in 2022 achieves the best distortion accuracy while subsequent methods, misled by the existing evaluation metrics, blindly improve the image quality while gradually degrading distortion accuracy. We assume that this phenomenon is caused by learning distortion and content completion in one model and the model tends to prioritize optimizing the latter.

In this work, we propose a decoupled panorama generation pipeline, termed PanoDecouple, aiming to generate panoramas with accurate distortion and visual appeal at once. Specifically, we decouple the task of panorama generation into distortion guidance and content completion. For distortion guidance, we first introduce a general panorama distortion representation, distortion map into the network as explicit distortion guidance, then we modify the condition registration mechanism of ControlNet from first block only to all the blocks to make it suitable for position-encoding-like condition input (*i.e.*, distortion map). As for content completion, we basically follow the architecture of mask-based outpainting methods [52] but substitute the text condition with perspective image information to better align the content of the generated panorama with the NFoV image. The two branches will be fused into the pre-trained, frozen U-Net, leveraging the powerful prior knowledge of pre-trained data and handling the information fusion simultaneously. What’s more, we design a distortion correction loss, utilizing the distortion prior in Distort-CLIP to further constrain the distortion.

By introducing the decoupled pipeline, we achieve the best image quality and second-best distortion accuracy in SUN360 and SOTA performance in Laval Indoor in zero-shot manner. Notably, we only use 3K training data, which is 15 times less than the existing methods, but achieved surprising generalization ability, highlighting the importance of decoupling the task of panorama generation. To further validate the universality of our PanoDecouple, we propose two practical applications (*i.e.*, text editing and text-to-image panorama generation) and show appealing results.

In summary, our main contributions are as follows:

(1) We identify the limitation in current evaluation metrics for assessing panorama distortion and introduce a distortion-aware model, Distort-CLIP and corresponding

evaluation metrics Distort-FID. Based on Distort-CLIP, we further observe the “visual cheating” phenomenon in existing methods and trace its cause to learn panorama distortion and content completion in a single model.

(2) A decoupled panorama generation pipeline, PanoDecouple is proposed. PanoDecouple introduces a novel distortion guidance branch, DistortNet into the network with an accurate distortion prior and a corresponding condition registration mechanism.

(3) Our PanoDecouple surpasses prior methods on two benchmarks with only 3K training data. PanoDecouple can also be applied to text editing and text-to-image panorama generation for free with appealing results.

2. Related Work

2.1. Image Outpainting

Image outpainting [4, 8, 10, 24, 35, 44, 50, 52] is a challenging computer vision task, and is an important step toward Artificial General Intelligence [21–23, 26, 46, 47, 49, 54–56], which requires expanding a partial image with content and style consistency. Unlike inpainting, outpainting struggles to utilize information from pixels near the unknown area. SRN [44] first leverages GAN [12] for outpainting with the guidance of incomplete semantic information; QueryOTR [50] splits the image into patches, considering it as queries and utilize vision transformer [7, 41] to outpaint the unknown patches assisted by queries. With the development of VQ-VAE [8, 16, 40], diffusion model [9, 14, 27, 31, 35, 37] and ControlNet [20, 52], mask-based diffusion outpainting methods apply extra conditions into ControlNet and achieve great performance by leveraging the powerful prior knowledge of latent diffusion.

Although the methods above can achieve great results in content completion, they are not suitable for panorama generation because of the characteristics of panorama generation, which consists of distortion guidance and content completion. In this work, We design a decoupled diffusion model, consisting of a DistortNet and a ContentNet to address the two characteristics respectively.

2.2. Panorama Generation

Panorama generation contains two settings: text-to-image generation and image outpainting. Text-to-image generation [3, 6, 17, 18, 39, 42, 51, 53] free the constrain of image input and fully utilize the prior knowledge of pre-trained latent diffusion model to generate fancy panorama. Text2Light [6] proposes distortion map, a general panorama distortion representation to constrain the generated distortion by considering it as the prompt tokens in the transformer. PanFusion [51] sets the distortion map as position encoding and utilizes its information in an attention mechanism. As for panoramic image outpainting [1, 25, 29, 43, 45], in this setting, there’s a greater emphasis on ensuring

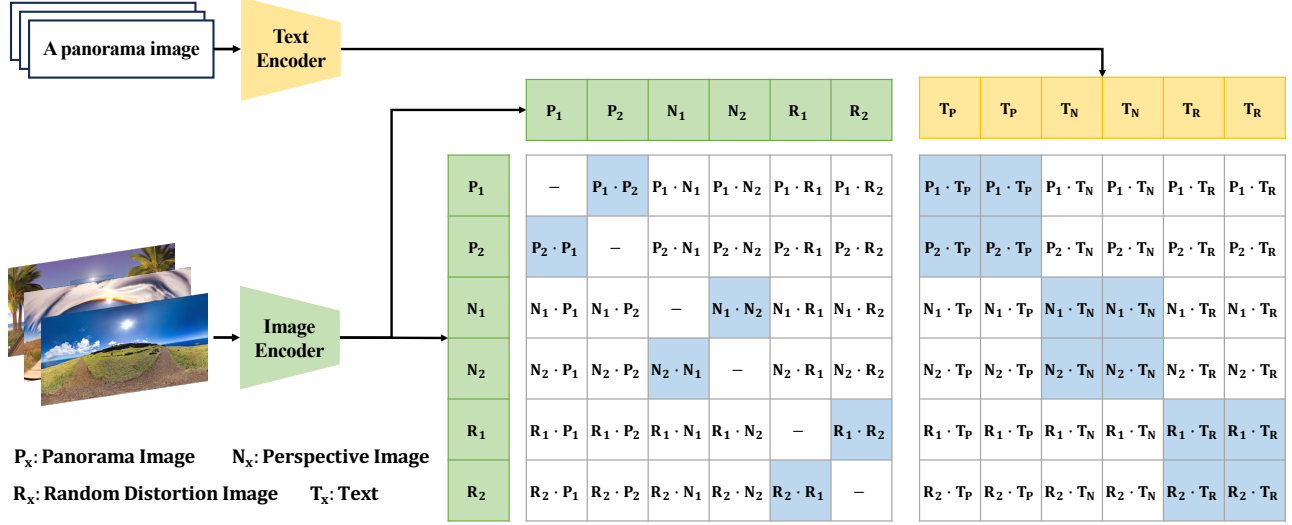


Figure 2. The training pipeline of our Distort-CLIP. The image features of three distortion types will do cosine similarity with themselves, and text features of three distortion types respectively. “-” means that the corresponding elements will not participate in the computation because it is meaningless. The boxes in blue mean the similarity of corresponding elements is 1, otherwise 0. Zoom in for best view.

the consistency of the generated panorama with the content and style of the given NFoV image, while also maintaining accurate panoramic distortion. However, existing methods mainly consider the image quality of the generated panorama, while ignoring the distortion and causing “visual cheating” phenomenon as shown in Fig 1.

In this work, we adopt the distortion map followed [6, 51] to constrain the distortion accuracy. However, we do not leverage it in attention mechanism, but modify the original ControlNet architecture and make it suitable for position-encoding-like condition input (*i.e.*, distortion map).

3. Distort-CLIP

In this section, we first show the motivation for our distortion analysis methods, then we generate data needed for evaluation and use the generated data to show the limited ability of existing metrics for evaluating distortion. Finally, we propose Distort-CLIP to focus on assessing the panorama distortion.

3.1. Existing Metrics Evaluation

Motivation. The best choice for evaluating the models’ distortion assessment ability (used in existing metrics) is comparing the similarity of image pairs of “different contents, same distortion” and “same contents, different distortion”. A good model will exhibit high feature similarity for image pairs with different content but the same distortion, and low similarity otherwise.

Data Generation. We aim to generate no-distortion (perspective) images that are consistent with the panoramic content to validate the model’s capability systematically. Specifically, we first extract the central region of panoramas data (*i.e.*, SUN360 [48]) and project it into perspective

Table 1. Comparison of our Distort-CLIP with other models used in evaluation metric. We show the feature similarity (range from -1 to 1) between different pairs (*i.e.*, different distortion, same content; same distortion, different content; panorama and different distortion texts). The best results are in **bold**.

Method	InceptionNet	CLIP	Distort-CLIP
Pano-Pers ↓	0.490	0.752	0.001
Pano _{train} -Pano _{test} ↑	0.229	0.693	0.993
Pano _{train} -Text _{pano} ↑	-	0.271	0.995
Pano _{train} -Text _{pers} ↓	-	0.269	0.001

images followed [51], then we apply sd-outpainting [52] model to outpaint the perspective images with the same text as panoramas but an extra prefix “A perspective image of”.

Results. We show the feature similarity between panoramas and perspective images with same content, panorama and panorama with different content in Table 1. The results show that the InceptionNet (*i.e.*, used in FID, IS) and CLIP (*i.e.*, used in CLIP-FID, CLIP-Score) **tend to perceive image content information**, with a fragile ability to perceive distortions, especially in the case of InceptionNet. So we claim that existing evaluation metrics used in panorama generation are not suitable for distortion assessment and a more robust model is needed.

3.2. Distort-CLIP Tuning

Based on the result in Table 1, CLIP owns a certain ability to perceive the distortion and we think that it benefits from the contrastive learning mechanism. So we fine-tune the CLIP with panoramas and our generated images in contrastive learning mechanism. Here we add an additional random distortion types images that are consistent with the panoramic content to improve the robustness of the

model. For the details of generating random distortion images, please refer to *Appendix A*. Since this model is used for metric evaluation and subsequent distortion correction loss, it doesn't require generalization. Therefore, we directly fine-tune the CLIP network as shown in Fig. 2. We show the details below.

Image Encoder. The image encoder is used to calculate the evaluation metric, which aims to distinguish the panoramic distortion in the image. So we input the three types of images (*i.e.*, normal, random distortion, panoramic images) into the network at the same time and the training objective is to optimize the cosine similarity between the image pairs with same distortion type while enlarging the difference with others. The contrastive loss is defined as follows:

$$\mathcal{L}_{ie}(x, y) = \frac{1}{N(N-1)} \sum_{i=1}^N \sum_{j=1, j \neq i}^N (x_i^\top y_j - l_{ij})^2, \quad (1)$$

where x_i , y_j are the normalized image vectors, l_{ij} means the label of this index (*i.e.*, 1 if x_i and y_j are the same type and 0 otherwise). We discard the self-similarity calculation (diagonal), as it holds no practical significance.

Text Encoder. The text encoder is utilized to compute the distortion correction loss function (see Sec 4.4 for details). This process requires merely distinguishing distortions rather than providing a detailed description of the image. Therefore, we only use three text prompts as input (*i.e.*, "A panorama image", "A perspective image" and "A random distortion image") to fine-tune the text encoder. The contrastive loss is defined as follows:

$$\mathcal{L}_{te}(x, z) = \frac{1}{N \times N} \sum_{i=1}^N \sum_{j=1}^N (x_i^\top z_j - l_{ij})^2, \quad (2)$$

where z_j is the normalized text vectors, the $l_{i,j}$ in \mathcal{L}_{te} is the same as the one in \mathcal{L}_{ie} as the distortion type of the image and text corresponds to each other. Additionally, recognizing the distinct roles of the two encoders, we cut off the gradient flow to the image encoder while tuning the text encoder. This strategy enables each encoder to achieve its full potential. The whole loss function is shown as follows:

$$\mathcal{L} = \mathcal{L}_{te} + \mathcal{L}_{ie}. \quad (3)$$

Results. As shown in Table 1, after fine-tuning the CLIP, we significantly improve the ability to distinguish the distortion both in image-image and image-text manners. Note that the test set of SUN360 (Pano_{test}) is not covered within the training data range, yet Distort-CLIP can still accurately determine that their distortion types are the same, showing the robustness of our method. We further show the distortion correction ability of our Distort-CLIP in the experiment 5.3.

4. PanoDecouple

To avoid the "visual cheating" phenomenon in previous work, we design PanoDecouple, decoupling the task of

panorama generation into distortion guidance (DistortNet) and content completion (ContentNet) as shown in Fig 3. We further propose the distortion correction loss based on the proposed Distort-CLIP to constrain the distortion explicitly.

4.1. Preliminary

Latent Diffusion. Latent diffusion is first introduced by [35] to address the high cost of operations in image space. This pipeline first transforms the image x into latent map z via an autoencoder \mathcal{E} and then trains a U-Net network ϵ_θ to learn the denoising ability. The training objective is defined as follows:

$$\mathcal{L} = \mathbb{E}_{\mathcal{E}(x), t, \epsilon, c_t} [\|\epsilon - \epsilon_\theta(z_t, t, c_t)\|_2^2], \quad (4)$$

where z_t means the noised latent map in the timestep t and c_t is the text condition. In the reverse process, z_T is initialized by standard Gaussian noise and is fed into ϵ_θ iteratively to obtain the clean latent map z_0 . Finally, the decoder \mathcal{D} is applied to decode the z_0 into image space.

ControlNet-based Panorama Outpainting. ControlNet [52] provides a way for diverse conditions control by adding a control branch, and the adding mechanism in mask-based panorama outpainting is as follows:

$$out = \mathcal{F}_m(z) + \mathcal{Z}(\mathcal{F}_{cn}(z, c_t, c_p, \mathcal{M})), \quad (5)$$

where \mathcal{F}_m is the main U-Net, \mathcal{F}_{cn} is the ControlNet which initializes by the parameters of U-Net encoder part. \mathcal{Z} is the zero convolution layer, which aims to avoid large modifications in the early stage. z , c_t , c_p , \mathcal{M} are the latent feature, text, partial panorama and the outpainting region mask respectively. **Note that conditions c_p and \mathcal{M} are only added in the first block in the source code of ControlNet and the others only have text condition input as follows:**

$$\mathcal{F}_{cn}(z, c_t, c_p, \mathcal{M}) : \begin{cases} cn^0 = CN(z, c_t) + CE(c_p, \mathcal{M}), \\ cn^b = CN(cn^{b-1}, c_t), \quad b \in (1, B), \end{cases} \quad (6)$$

where CE and CN are the content encoder and ControlNet similar to ours in Fig 3, B is the block number.

4.2. DistortNet

Panorama is a unit sphere in 3D space, which is also the resource of the distortion. Therefore, a reasonable choice for achieving distortion guidance is searching for an explicit distortion representation, which allows the model to perceive the relationship between corresponding points in 2D image space and 3D space. We introduce a suitable representation, distortion map below and then design a DistortNet to fully leverage the potential of the distortion map.

Distortion Map Construction. The distortion map D is built followed [6, 51], which is a 2D position signals for 3D scene representation S (*i.e.*, panorama is a unit sphere in 3D space). Given a panoramic image $I \in \mathbb{R}^{H \times W \times 3}$, the pixel relationship with sphere S is defined as follows:

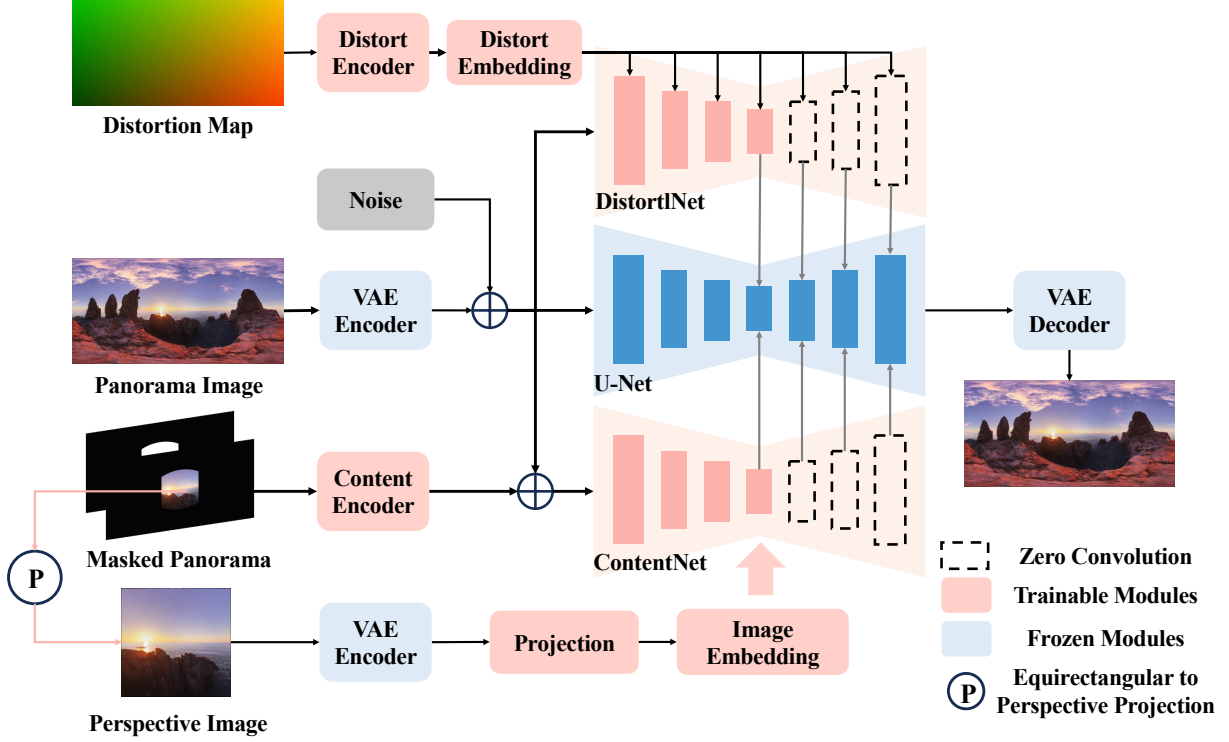


Figure 3. The pipeline of the proposed PanoDecouple, a decoupled diffusion model. The DistortNet focuses on distortion guidance via the proposed distortion map. To make full use of the position-encoding-like distortion map, we modify the condition registration mechanism of ControlNet from the first block only to all the blocks. The ContentNet is devoted to content completion by imposing partial panorama image input and perspective information. The U-Net remains frozen, coordinating the information fusion between content completion and distortion guidance branches, while fully leveraging its powerful pre-trained knowledge. Note that we omit the text input of the DistortNet and U-Net for simplification while the one for ContentNet is replaced by perspective image embedding.

$$\begin{aligned}
 S(\theta, \phi, r) &= I(i, j), \\
 \theta &= (2i/H - 1)\pi, \\
 \phi &= (2j/W - 1)\pi/2, \\
 r &= 1,
 \end{aligned} \tag{7}$$

where i, j is the pixel in panorama, θ and ϕ represent the azimuth angle (range from $-\pi$ to π) and the elevation angle (range from $-\pi/2$ to $\pi/2$) of the sphere, respectively. The center of the sphere S is the position of the camera. So the distortion map $D \in \mathbb{R}^{H \times W \times 2}$ is defined as follows:

$$D(i, j) = (\theta, \phi). \tag{8}$$

The formula defines the correspondence between a pixel on a panorama image and a point on a 3D sphere. What's more, since a panorama is a sphere in 3D space, its left and right edges can seamlessly align in a 2D representation. However, the current distortion map D does not exhibit this property (*i.e.*, $-\pi$ and π in the leftmost and rightmost edge). Therefore, we introduce first-order Taylor expansion positional encoding [28] $\gamma(\cdot)$ to make it continuous as follows:

$$\begin{aligned}
 D(i, j) &= (\gamma(\theta), \gamma(\phi)), \\
 \gamma(\theta) &= [\sin(2^0 \pi \theta), \cos(2^0 \pi \theta)], \\
 \gamma(\phi) &= [\sin(2^0 \pi \phi), \cos(2^0 \pi \phi)].
 \end{aligned} \tag{9}$$

In this way, the edges are continuous ($\sin(-\pi)$ and $\sin(\pi)$, $\cos(-\pi)$ and $\cos(\pi)$ are equal). The obtained $D \in \mathbb{R}^{H \times W \times 4}$ will be fed into DistortNet and we name it c_d for uniform.

Condition Registration Mechanism Modification. As mentioned in Sec 4.1 Eq. 6, this operation benefits condition types like image, depth, pose but not position-encoding-like distortion map as commonly the position encoding should interact with model in each block (*e.g.*, timestep t in diffusion model [14], position embedding in ViT [7]).

In this work, we modify the condition registration mechanism from the first block only to all the blocks as follows:

$$\mathcal{F}_{dn}(z, c_t, c_d) : \begin{cases} dn^b = DN(dn^{b-1}, c_t) + \mathcal{Z}(Proj^b(de)), \\ de = DE(c_d), \end{cases} \tag{10}$$

where DN , DE and dn are the ‘‘DistortNet’’, ‘‘Distort Encoder’’ and ‘‘Distort Embedding’’ in Fig 3 respectively, $Proj^b$ is the 2D convolution to map the de into dimension in b^{th} block. We further compare our modification with attention and original ControlNet mechanism in experiment.

4.3. Decoupled Diffusion Model

ContentNet. The ContentNet basically follows the architecture of previous mask-based outpainting method [43].

Table 2. Comparison with SOTA methods. † means re-implementing in our setting for fair comparison. Note that the bottom region of Laval is entirely black edges and we crop 20% of it when testing image quality and undo it when testing distortion as it requires full image. (·) means the crop setting of PanoDiff (crop 20% up and bottom region). The best, second-best results are in **bold**, underline.

Method	Year	Training samples	SUN360				Laval Indoor			
			FID ↓	CLIP-FID ↓	Distort-FID ↓	IS ↑	FID ↓	CLIP-FID ↓	Distort-FID ↓	IS ↑
OmniDreamer	2022	50K	75.14	11.13	0.52	4.58	133.15 (118.8)	22.47 (11.21)	<u>2.95</u>	2.82 (3.04)
ImmerseGAN	2022	250K	-	9.26	-	-	-	-(12.69)	-	-
PanoDiff †	2023	3K	<u>63.49</u>	<u>7.04</u>	2.68	<u>6.51</u>	<u>80.63</u> (<u>77.25</u>)	<u>10.26</u> (<u>7.77</u>)	13.53	<u>2.90</u> (<u>3.17</u>)
AOG-Net †	2024	3K	74.07	8.87	4.52	6.32	93.09 (89.38)	12.04 (9.25)	11.59	2.65 (2.89)
2S-ODIS †	2024	3K+9K	72.23	9.11	8.23	5.56	81.43 (80.67)	<u>9.45</u> (7.97)	10.21	2.54 (2.60)
PanoDecouple (Ours)	-	3K	62.19	6.21	<u>0.92</u>	6.93	74.45 (72.51)	8.88 (7.30)	2.85	2.95 (3.28)

However, we substitute the text embedding with the proposed perspective image embedding to avoid the text-image inconsistent problem [5, 9, 33] caused by BLIP-2 [19]. Specifically, we use the Equirectangular-to-Perspective Projection as mentioned in *Appendix* of [51] to project the central region of panoramic image c_p into the perspective image c_n as shown in Fig 3. Then the “VAE Encoder” and “Projection” networks are used to match the dimension with original text embedding as follows:

$$c'_n = Proj(VAE(c_n)). \quad (11)$$

Fusion in U-Net. The functions of the main U-Net are leveraging the powerful prior knowledge and coordinating the information of distortion guidance and content completion branches simultaneously. The fusion pipeline of each block is defined as:

$$out = \mathcal{F}_m(z) + \mathcal{Z}(\mathcal{F}_{cn}(z, c'_n, c_p, \mathcal{M})) + \mathcal{Z}(\mathcal{F}_{dn}(z, c_t, c_d)), \quad (12)$$

4.4. Distortion Correction Loss

To generate more accurate distortion, we propose the distortion correction loss function via the text encoder of the Distort-CLIP to constrain the generation output explicitly. Specifically, we first transform the model output ϵ into clean latent code z_0 by leveraging the forward diffusion process:

$$z_0 = \frac{1}{\sqrt{\alpha_t}}(z_t - \sqrt{1 - \alpha_t}\epsilon), \quad (13)$$

where z_t is the noised latent code in timestep t and α is the noise coefficient. Then we use the VAE decoder \mathcal{D} to decode z_0 into image space x as follows:

$$x = \mathcal{D}(z_0). \quad (14)$$

After obtaining the generated panoramic image x , we compute its cosine similarity with the text of three types of distortion T_P , T_N and T_R , which is similar to Fig 2. We set x , z_P , z_N and z_R as the normalized image and text feature vectors after Distort-CLIP encoders for simplification. The loss function is defined as follows:

$$\mathcal{L}_{dist} = x^T z_P - x^T z_N - x^T z_R. \quad (15)$$

The full loss function for training our PanoDecouple is as follows:

$$\mathcal{L} = \mathcal{L}_{rec} - \lambda \cdot \mathcal{L}_{dist}, \quad (16)$$

where \mathcal{L}_{rec} is the same as Eq. 4 and λ is the coefficient which we set at 0.05 in the experiment.

5. Experiments

5.1. Experimental Setup

Datasets. We follow PanoDiff [43] to conduct experiments in SUN360 [48] and Laval Indoor [11] datasets. SUN360 comprises both indoor and outdoor scenes while Laval Indoor only has indoor scene. We use 3000/500 SUN360 data for training/testing and 289 Laval Indoor data for zero-shot generalization. **Note** that it is not the full number of SUN360 as the official link is unavailable and for the methods with a larger training number than ours, we directly test their pre-trained models on our test set to maintain fair comparison. As for the methods which does not open the pre-trained model, we re-implement them in our setting. The text prompts used in our methods are generated with BLIP-2 [19]

Evaluation Metrics. As previous evaluation metrics exist problems in perceiving distortion. We utilize Fréchet Inception Distance (FID) [13], CLIP-FID and Inception Score (IS) [36] to evaluate the panoramic image quality and use the proposed Distort-CLIP to compute the Distort-FID followed the calculation process of FID.

Implementation Details. PanoDecouple is trained based on the pre-trained weight of [52], using one H20 GPU with PyTorch [30] environment. During training, We use the Adam [15] optimizer and L2 loss for 20 epochs with the initial learning rate 1e-5 and batch sizes 2. During inference, we use 50 timesteps for all the experiments.

5.2. Evaluation on the benchmarks

Quantitative Results. The result is shown in Table 2, in SUN360, the OmniDreamer [1] achieves best distortion accuracy with limited image quality and subsequent methods, constrained by evaluation metrics, blindly improved image quality, gradually ruining distortion accuracy, which is the “visual cheating” phenomenon we pointed out. Our PanoDecouple significantly reduces the distortion inaccuracy



Figure 4. Qualitative comparison of panorama generation from N FoV image. We sequentially present the results on SUN360, Laval Indoor, and raw images (two images each). Zoom in for best view.

in recent methods and achieves great image quality at the same time. In Laval Indoor, as the ceiling is hard to generate accurately, PanoDiff crops it while we maintain it as a harder setting. Notably, by using only 3K training data, we could achieve state-of-the-art image quality and great distortion accuracy, especially in Laval Indoor (zero-shot manner), showing that decoupling the panorama generation into distortion guidance and content completion truly benefits performance improvement and generalization ability.

Qualitative Results. The result is shown in Fig 4 (*i.e.*, Appendix C for perspective and more panorama results), both in SUN360, Laval Indoor and raw images, we achieve great image quality and distortion accuracy at once. Notably, we only use 3K training data, but showing great generalization ability even in raw website images. This phenomenon validates the significance of decoupling panorama generation into distortion guidance and content completion.

5.3. Ablation Study

We conduct ablation study on SUN360. We first evaluate the effectiveness of each component we proposed and validate the synergistic effect; then we explore the best condition registration mechanism for position-encoding-like distortion map; finally we prove the effectiveness of using general distortion representation, distortion map.

Table 3. Quantitative comparison for the ablation study. SD, MD, PN, DLoss mean first-block condition registration in original ControlNet, all-block condition registration in our DistortNet, perspective image embedding and distortion correction loss respectively. The best, second-best results are in **bold** and underline.

Method	SD	MD	PN	DLoss	CLIP-FID ↓	Distort-FID ↓	IS ↑
Ours	✓				7.04	2.68	6.51
		✓			6.66	1.63	6.16
		✓	✓		<u>6.38</u>	1.05	6.35
		✓	✓		6.45	<u>0.98</u>	6.95
		✓	✓	✓	6.21	0.92	<u>6.93</u>

The Synergistic Effect Analysis. We evaluate the effectiveness of the proposed DistortNet, ContentNet and distortion correction loss in Table 3. Note that without any modules we proposed is an image-outpainting diffusion model (PanoDiff [43]). By feeding the distortion map into extra ControlNet in first-block condition registration mechanism, we significantly improve the distortion accuracy while degrading the generation diversity (IS) as adding additional condition naturally affects the diversity of network generation. When modifying the mechanism into all-block registration (*i.e.*, our proposed DistortNet), we further improve the distortion accuracy. It could be seen that benefiting from decoupling the network, we ease the task of content com-



‘The dragon in Game of Thrones’



‘Pikachu is using the move Thunderbolt’

Figure 5. The quantitative results of text-panorama generation. Zoom in for best view.

pletion and the image quality is gradually improved. When replacing the text embedding in ContentNet with perspective image embedding, we largely improve the generation diversity, benefiting from consistent representation between panoramas and N FoV images. Finally, by introducing the distortion correction loss into the model, not only the distortion accuracy is refined, but also the decoupling function is enhanced and simplifies the task of content completion, improving the image quality. In summary, the design of the decoupling diffusion model achieves a synergistic effect in both image quality and distortion constraint.

Condition Registration Mechanism. Here we systematically analyze which condition registration mechanism is suitable for position-encoding-like distortion map. We design four model architectures for analysis, the detailed architecture is shown in *Appendix B* and the results are shown in Table 4, utilizing all-block condition registration mechanism works better than the attention and original ControlNet mechanism. We hope this observation will contribute to advancing future research on diverse condition injection in diffusion models.

Distortion Condition Selection. To validate the effectiveness of distortion map, we replace it with full-zero map, learnable map and distortion map without first-order Taylor extension as shown in Table 4, the learnable map achieves certain improvements while degrading the image diversity as learning the distortion map complicates the task. The introduced distortion map achieves the best result as it offers the network prior knowledge about panorama distortion representation, while without Taylor extension, the leftmost and rightmost consistency is ruined.

5.4. Applications

Text Editing. With different text prompts and the same N FoV image, our PanoDecouple can generate panoramas with different scenes and styles as shown in *Appendix*.

Random N FoV Image Input. PanoDecouple could achieve it by few step tuning based on original 90-degree central N FoV models. We show the result in *Appendix*.

Text-to-Panorama Generation. Our PanoDecouple can freely apply to another setting: text-to-panorama generation by first utilizing a pre-trained text-to-image diffusion

Table 4. Quantitative comparison for the ablation study about the condition registration mechanism and distortion condition selection. The best results are in **bold**.

Method	FID ↓	CLIP-FID ↓	Distort-FID ↓	IS ↑
<i>Condition Registration Mechanism</i>				
attn-unet	65.34	6.97	2.52	6.06
add-contorlnet	62.95	6.54	1.59	6.83
attn-contorlnet	65.71	7.92	5.22	5.58
attn-distortnet	63.54	6.56	1.48	6.57
add-distortnet	62.19	6.21	0.92	6.93
<i>Distortion Condition Selection</i>				
zero map	63.96	6.57	2.40	6.47
learnable map	63.32	6.43	1.42	6.23
distortion w/o Taylor	62.57	6.68	2.48	5.78
distortion map	62.19	6.21	0.92	6.93

model, SDXL [32] and then feeding the generated image into our PanoDecouple. The result is shown in Fig 5 bottom line. We generate panoramas with visual appeal, accurate distortion and text consistency simultaneously.

6. Conclusion

In this work, we experimentally discover the limitation of existing evaluation metrics in perceiving panorama distortion and propose a distortion-specific Distort-CLIP to address it. Based on Distort-CLIP, we further observe the visual cheating phenomenon in previous works which tends to improve the image quality while sacrificing the distortion accuracy. We think that this is caused by learning panorama distortion and content completion in a single model at once and propose a decoupled panorama generation pipeline, PanoDecouple, to address it. PanoDecouple consists of a DistortNet by imposing panorama distortion prior and a corresponding condition registration mechanism; and a ContentNet equipped with natural image information. Extensive experiments show that PanoDecouple achieves accurate distortion and visual appeal at once.

Acknowledgments. This work was supported partially by the National Key Research and Development Program of China (2023YFA1008503), NSFC(92470202, U21A20471), Guangdong NSF Project (No. 2023B1515040025); and Alibaba Group through Alibaba Research Intern Program.

References

- [1] Naofumi Akimoto, Yuhi Matsuo, and Yoshimitsu Aoki. Diverse plausible 360-degree image outpainting for efficient 3dgc background creation. In *CVPR*, 2022. 2, 6
- [2] Jiayang Bai, Letian Huang, Jie Guo, Wen Gong, Yuanqi Li, and Yanwen Guo. 360-gs: Layout-guided panoramic gaussian splatting for indoor roaming. *arXiv preprint arXiv:2402.00763*, 2024. 1
- [3] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. In *ICML*, 2023. 2
- [4] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *CVPR*, 2022. 2
- [5] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart- α : Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023. 6
- [6] Zhaoxi Chen, Guangcong Wang, and Ziwei Liu. Text2light: Zero-shot text-driven hdr panorama generation. *TOG*, 2022. 2, 3, 4
- [7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 5
- [8] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021. 2
- [9] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024. 2, 6
- [10] Przemek Gardias, Eric Arthur, and Huaming Sun. Enhanced residual networks for context-based image outpainting. *arXiv preprint arXiv:2005.06723*, 2020. 2
- [11] Marc-André Gardner, Kalyan Sunkavalli, Ersin Yumer, Xiaohui Shen, Emiliano Gambaretto, Christian Gagné, and Jean-François Lalonde. Learning to predict indoor illumination from a single image. *arXiv preprint arXiv:1704.00090*, 2017. 6
- [12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 2
- [13] 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. In *NeurIPS*, 2017. 6
- [14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 2, 5
- [15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6
- [16] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 2
- [17] Yuseung Lee, Kunho Kim, Hyunjin Kim, and Minhyuk Sung. Syncdiffusion: Coherent montage via synchronized joint diffusions. In *NeurIPS*, 2023. 2
- [18] Jialu Li and Mohit Bansal. Panogen: Text-conditioned panoramic environment generation for vision-and-language navigation. In *NeurIPS*, 2023. 2
- [19] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 6
- [20] Ming Li, Taojiannan Yang, Huafeng Kuang, Jie Wu, Zhaoning Wang, Xuefeng Xiao, and Chen Chen. Controlnet++: Improving conditional controls with efficient consistency feedback. In *ECCV*, 2025. 2
- [21] Kun-Yu Lin, Jia-Run Du, Yipeng Gao, Jiaming Zhou, and Wei-Shi Zheng. Diversifying spatial-temporal perception for video domain generalization. In *NeurIPS*, 2023. 2
- [22] Kun-Yu Lin, Jiaming Zhou, and Wei-Shi Zheng. Human-centric transformer for domain adaptive action recognition. *TPAMI*, 2024.
- [23] Zuhao Liu, Xiao-Ming Wu, Dian Zheng, Kun-Yu Lin, and Wei-Shi Zheng. Generating anomalies for video anomaly detection with prompt-based feature mapping. In *CVPR*, 2023. 2
- [24] Chia-Ni Lu, Ya-Chu Chang, and Wei-Chen Chiu. Bridging the visual gap: Wide-range image blending. In *CVPR*, 2021. 2
- [25] Zhuqiang Lu, Kun Hu, Chaoyue Wang, Lei Bai, and Zhiyong Wang. Autoregressive omni-aware outpainting for open-vocabulary 360-degree image generation. In *AAAI*, 2024. 2
- [26] Zhen Lv, Yangqi Long, Congzhentao Huang, Cao Li, Chengfei Lv, Hao Ren, and Dian Zheng. Spatialdreamer: Self-supervised stereo video synthesis from monocular input. *arXiv preprint arXiv:2411.11934*, 2024. 2
- [27] Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. *arXiv preprint arXiv:2401.08740*, 2024. 2
- [28] B Mildenhall, PP Srinivasan, M Tancik, JT Barron, R Ramamoorthi, and R Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 5
- [29] Atsuya Nakata and Takao Yamanaka. 2s-odis: Two-stage omni-directional image synthesis by geometric distortion correction. *arXiv preprint arXiv:2409.09969*, 2024. 2
- [30] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 6
- [31] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023. 2
- [32] 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. *arXiv preprint arXiv:2307.01952*, 2023. 8, 11
- [33] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024. 6
- [34] 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 *ICML*, 2021. 1
- [35] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 4
- [36] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *NeurIPS*, 2016. 6
- [37] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2020. 2
- [38] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016. 1
- [39] Shitao Tang, Fuyang Zhang, Jiacheng Chen, Peng Wang, and Yasutaka Furukawa. Mvdiffrusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. *arXiv*, 2023. 2
- [40] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *NeurIPS*, 2017. 2
- [41] A Vaswani. Attention is all you need. In *NeurIPS*, 2017. 2
- [42] Hai Wang, Xiaoyu Xiang, Yuchen Fan, and Jing-Hao Xue. Customizing 360-degree panoramas through text-to-image diffusion models. In *WACV*, 2024. 2
- [43] Jionghao Wang, Ziyu Chen, Jun Ling, Rong Xie, and Li Song. 360-degree panorama generation from few unregistered nfov images. In *ACMMM*, 2023. 2, 5, 6, 7
- [44] Yi Wang, Xin Tao, Xiaoyong Shen, and Jiaya Jia. Wide-context semantic image extrapolation. In *CVPR*, 2019. 2
- [45] Tianhao Wu, Chuanxia Zheng, and Tat-Jen Cham. Panodiffusion: 360-degree panorama outpainting via diffusion. In *ICLR*, 2023. 2
- [46] Xiao-Ming Wu, Dian Zheng, Zuhao Liu, and Wei-Shi Zheng. Estimator meets equilibrium perspective: A rectified straight through estimator for binary neural networks training. In *ICCV*, 2023. 2
- [47] Xiao-Ming Wu, Jia-Feng Cai, Jian-Jian Jiang, Dian Zheng, Yi-Lin Wei, and Wei-Shi Zheng. An economic framework for 6-dof grasp detection. In *ECCV*, 2024. 2
- [48] Jianxiong Xiao, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Recognizing scene viewpoint using panoramic place representation. In *CVPR*, 2012. 3, 6
- [49] Guo-Hao Xu, Yi-Lin Wei, Dian Zheng, Xiao-Ming Wu, and Wei-Shi Zheng. Dexterous grasp transformer. In *CVPR*, 2024. 2
- [50] Kai Yao, Penglei Gao, Xi Yang, Jie Sun, Rui Zhang, and Kaizhu Huang. Outpainting by queries. In *ECCV*, 2022. 2
- [51] Cheng Zhang, Qianyi Wu, Camilo Cruz Gambardella, Xiaoshui Huang, Dinh Phung, Wanli Ouyang, and Jianfei Cai. Taming stable diffusion for text to 360 panorama image generation. In *CVPR*, 2024. 2, 3, 4, 6
- [52] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 2, 3, 4, 6
- [53] Qinsheng Zhang, Jiaming Song, Xun Huang, Yongxin Chen, and Ming-Yu Liu. Diffcollage: Parallel generation of large content with diffusion models. In *CVPR*, 2023. 2
- [54] Dian Zheng, Xiao-Ming Wu, Shuzhou Yang, Jian Zhang, Jian-Fang Hu, and Wei-Shi Zheng. Selective hourglass mapping for universal image restoration based on diffusion model. In *CVPR*, 2024. 2
- [55] Dian Zheng, Xiao-Ming Wu, Zuhao Liu, Jingke Meng, and Wei-shi Zheng. Diffuvolume: Diffusion model for volume based stereo matching. *IJCV*, 2025.
- [56] Jiaming Zhou, Kun-Yu Lin, Haoxin Li, and Wei-Shi Zheng. Graph-based high-order relation modeling for long-term action recognition. In *CVPR*, 2021. 2
- [57] Shijie Zhou, Zhiwen Fan, Dejie Xu, Haoran Chang, Pradyumna Chari, Tejas Bharadwaj, Suyu You, Zhangyang Wang, and Achuta Kadambi. Dreamscene360: Unconstrained text-to-3d scene generation with panoramic gaussian splatting. In *ECCV*, 2025. 1

Panorama Generation From NFoV Image Done Right

Supplementary Material



Figure S6. Visual results of panoramic images and corresponding perspective images. Zoom in for best view.

G. Random Distortion Image Generation

Motivation. The random distortion image could accelerate the robustness of our Distort-CLIP to assess distortion. The challenge is how to generate it while sharing the same content with panorama. We observe that the training data of stable diffusion [32] contains panorama image (with the height-width ratio 1:2). However, due to its relatively small proportion in training data, the distortion of the generated panoramic image is quite chaotic, which perfectly meets our requirements.

Method. We apply a mask to the panoramic image, leaving only the central area (*i.e.*, the same mask as shown in the main paper Fig. 3). Then we use the sd-outpainting based on [32] to outpaint the model with an extra prefix “A panorama image of”. In this way, we obtain random distortion images with the same content as panoramas.

Visual results. Here we also show the panoramic, generated perspective and random distortion images in Fig S6. It could be seen that we generate different distortion types images with similar content, ensuring the success of our Distort-CLIP.

H. Condition Registration Mechanism

We show the model architecture of different condition registration mechanisms in Fig S7. We explore the widely used mechanism for condition registration and validate that the all-block registration with addition is the best choice for position-encoding-like conditions.

I. More Applications

Text Editing. We show the text editing results below. Note that we specifically selected this NFoV image (with snow at

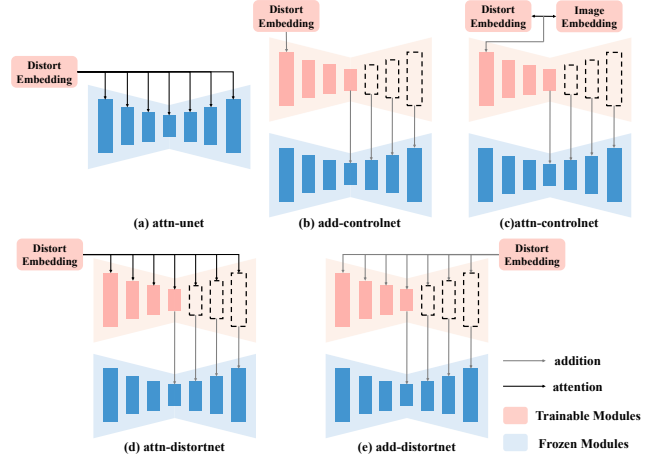


Figure S7. Various condition registration mechanisms. Note that both addition and attention own small learnable parameters (*e.g.*, projection layers and zero convolution for addition; common attention modules for attention), we omit it in figure for simplification. Zoom in for best view.

the bottom), which is extremely challenging for text editing when the text is inconsistent with the snow. However, our method still generates panoramas consistent with the text.



Figure S8. Visual results of text editing.

Random NFoV. We show results below, even the input image is extremely small and the position is tricky, we could generate panoramas with visual appealing.

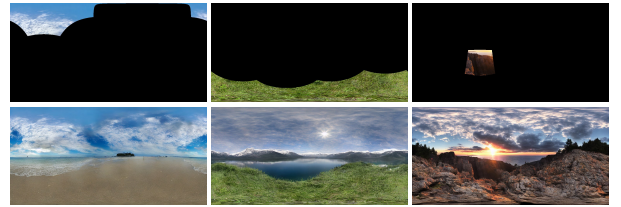


Figure S9. Visual results with random NFoV input.

J. More Visual Results

In Fig S10, we show panoramic and perspective results of generated panoramas by different methods. Our PanoDecouple achieves great image quality and accurate distortion simultaneously. We also show more raw image panorama outpainting results in Fig S11. Enjoy it!

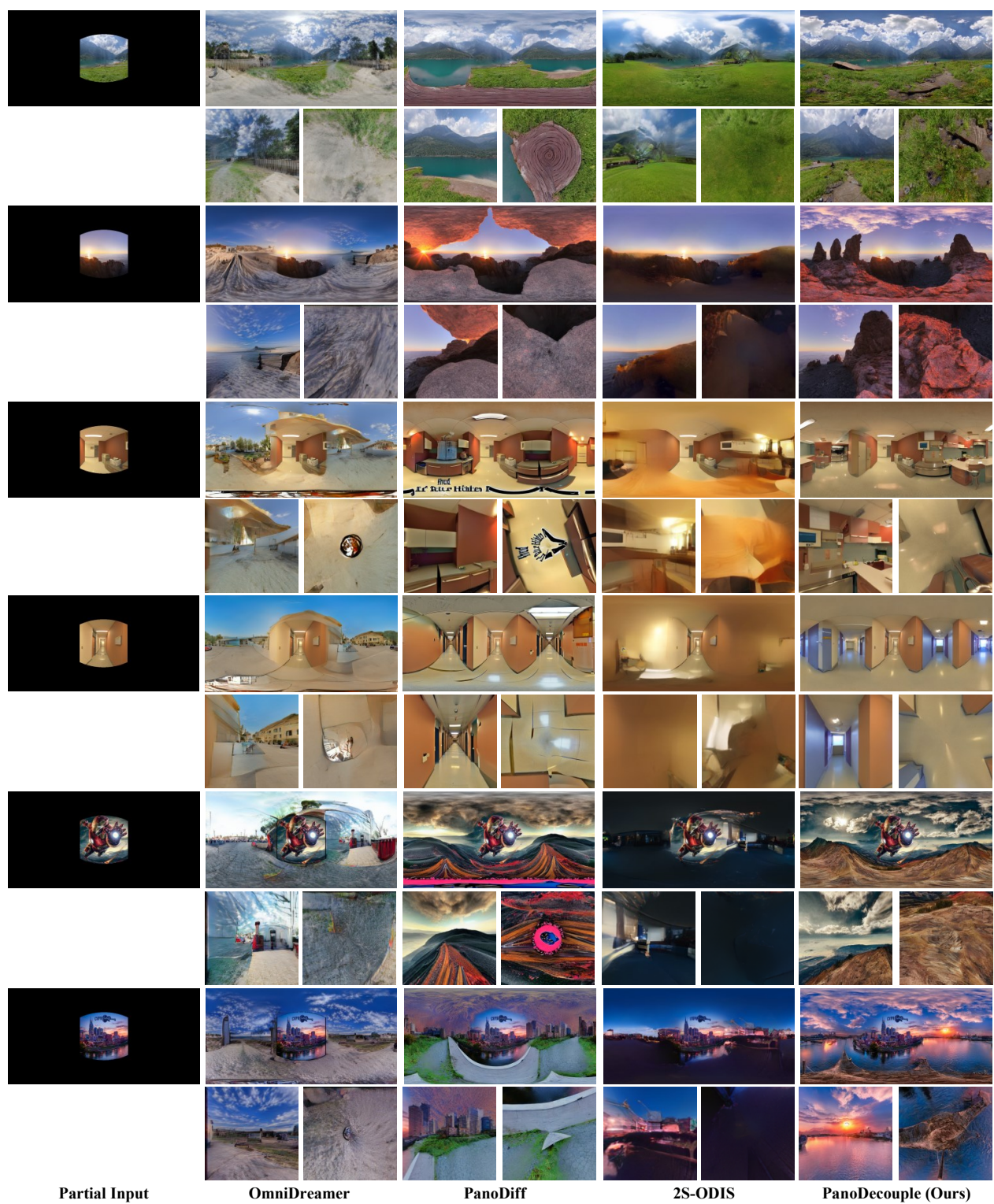
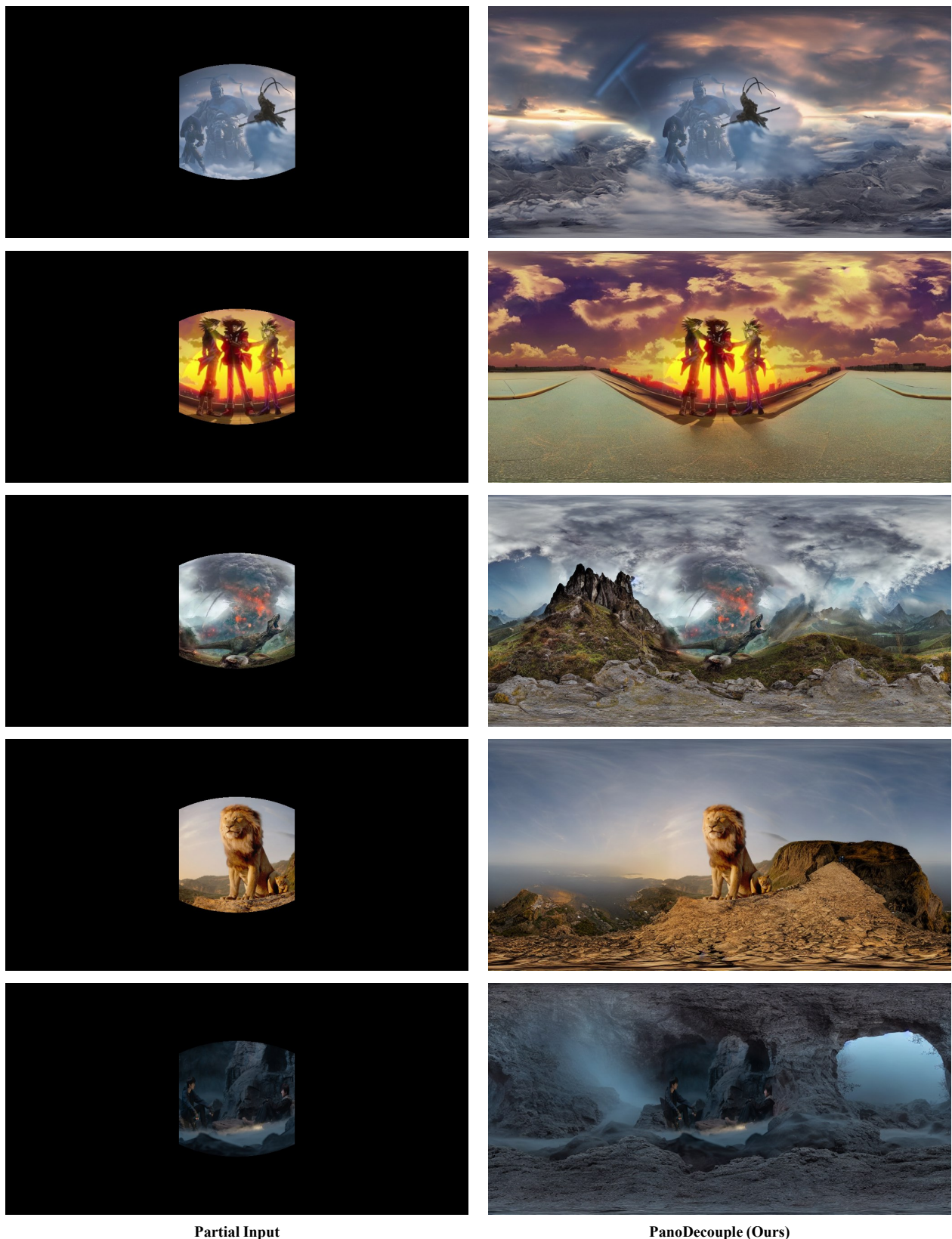


Figure S10. Visual results of panoramic images and corresponding perspective images.



Partial Input

PanoDecouple (Ours)

Figure S11. Visual results with raw image input. **Note that the images we use are for academic purposes only. If any copyright infringement occurs, we will promptly remove them.**