SyncDiffusion: Coherent Montage via Synchronized Joint Diffusions

"A photo of a city skyline at night"

NeurIPS 2023

Yuseung Lee Kunho Kim Hyunjin Kim Minhyuk Sung

Text-to-Image Diffusion Models

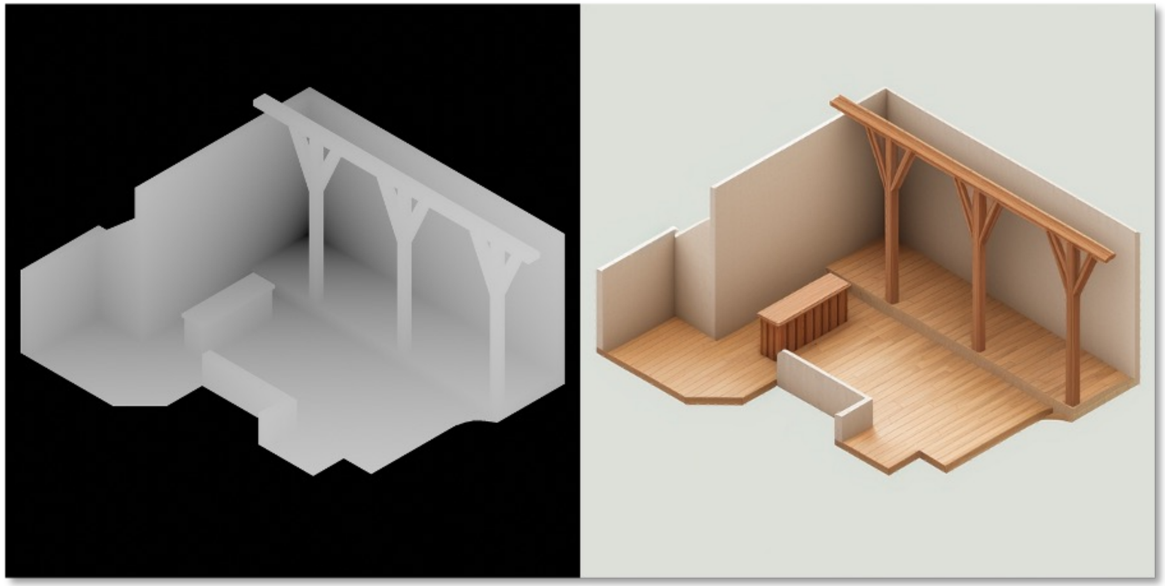
Pretrained text-to-image diffusion models are limited to generating images of certain sizes.

Stable Diffusion (Stability AI)

Needs for Arbitrary-Size Generation

There are growing demands for generating arbitrary-size images in downstream applications such as Virtual Reality and texture generation.

Virtual Reality (VR) Environment[1]

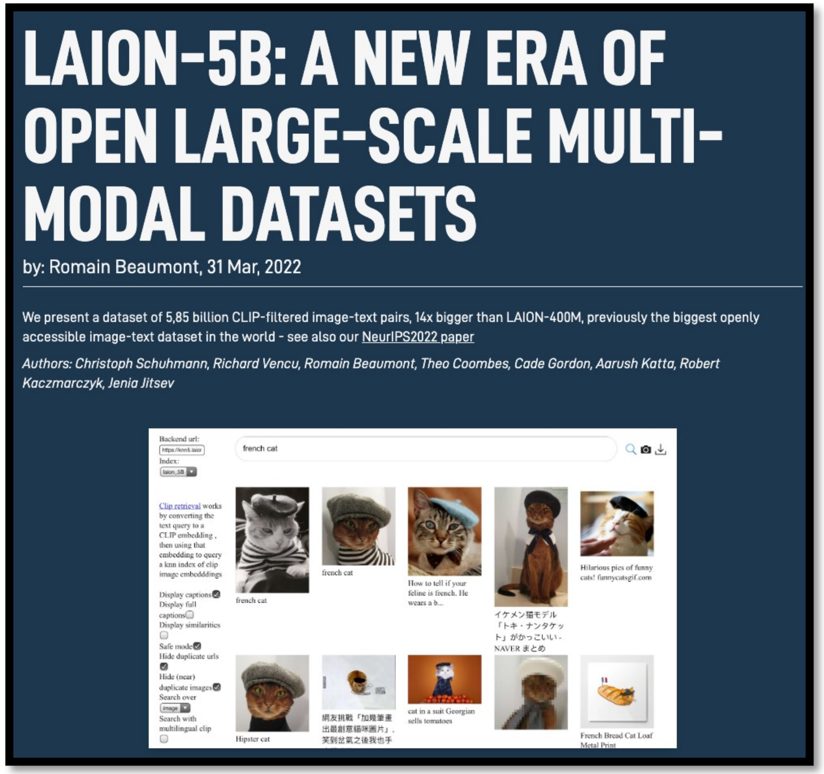


Generating textures for 3D objects[2]

[1] https://brdg.co/vr-room/
[2] https://medium.com/@mattzq/stable-diffusion-controlnet-texture-projection-workflow-with-blender-8dac4b7154c2

Expensive Data Acquisition & Training

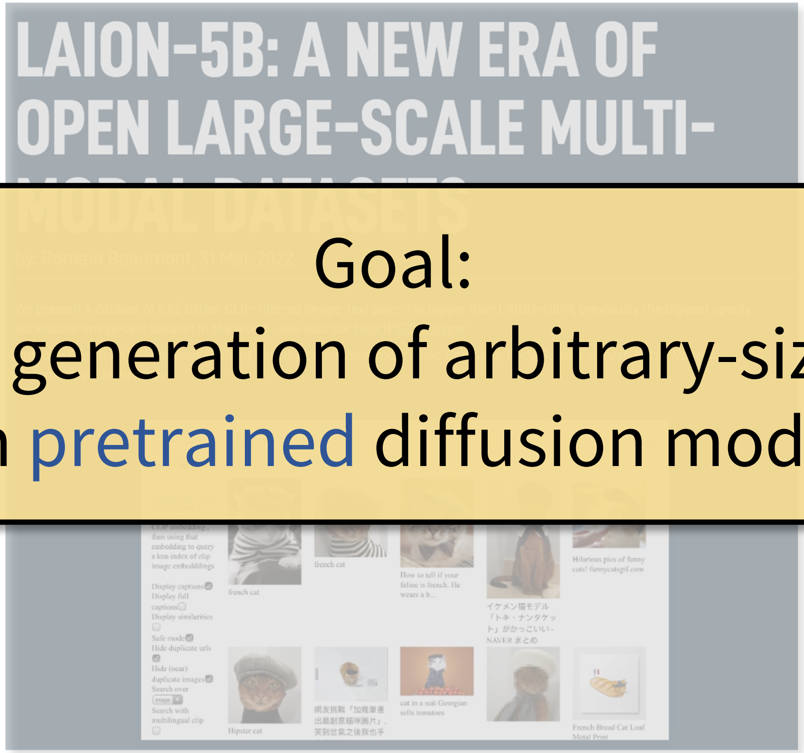
Training diffusion models for different image sizes would cost substantial time and computing resources.



LAION-5B

Expensive Data Acquisition & Training

Training diffusion models for different image sizes would cost substantial time and computing resources.



LAION-5B

Goal:

Zero-shot generation of arbitrary-size images with pretrained diffusion models.

Image as Montage

Any arbitrary-size image is a composition of multiple fixed-size images.

"A photo of a mountain range at twilight"

Image as Montage

Fixed-size images can be generated with pretrained models.

"A photo of a mountain range at twilight"

Image Extrapolation [Blended Latent Diffusion, Avrahami *et al.*]

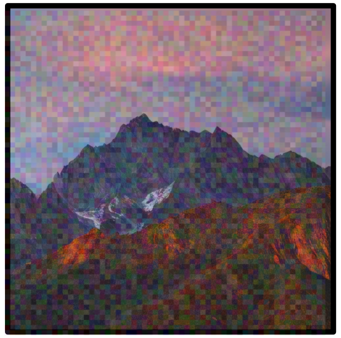
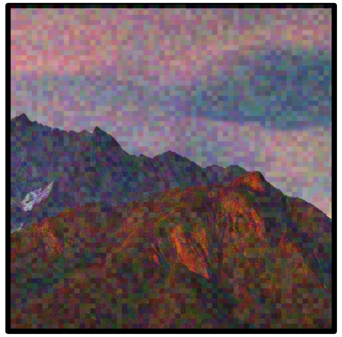
Sequentially extrapolating images often results in visible seams and repetitive contents.

Blended Latent Diffusion, Avrahami et al., SIGGRAPH 2023.

Joint Diffusion [MultiDiffusion, Bar-Tal *et al.*]

Average noisy latent features in overlapping regions.

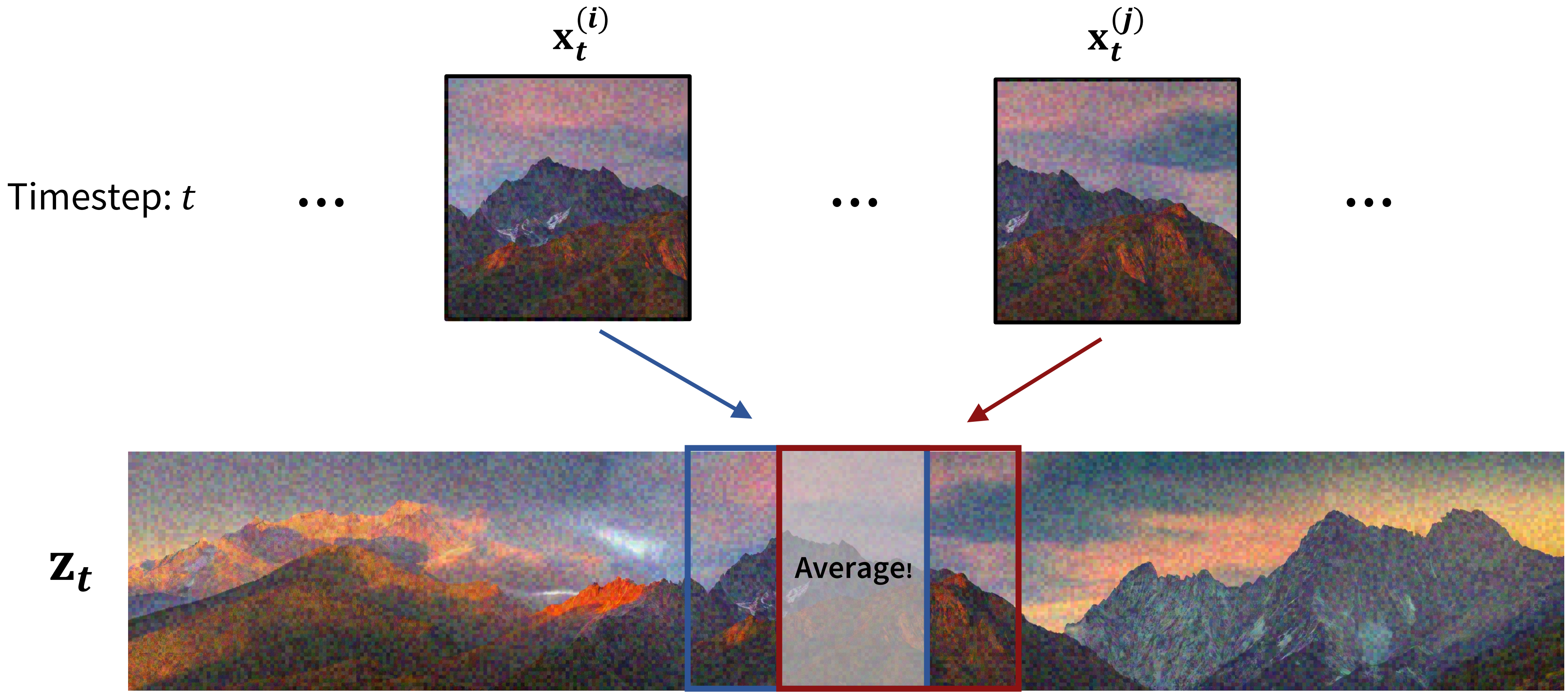
$\mathbf{x}_t^{(i)}$ $\mathbf{x}_t^{(j)}$

Timestep: t …  …  …

Bar-Tal et al., MultiDiffusion: Fusing Diffusion Paths for Controlled Image Generation, ICML 2023.

Joint Diffusion [MultiDiffusion, Bar-Tal *et al.*]

Average noisy latent features in the overlapping regions.

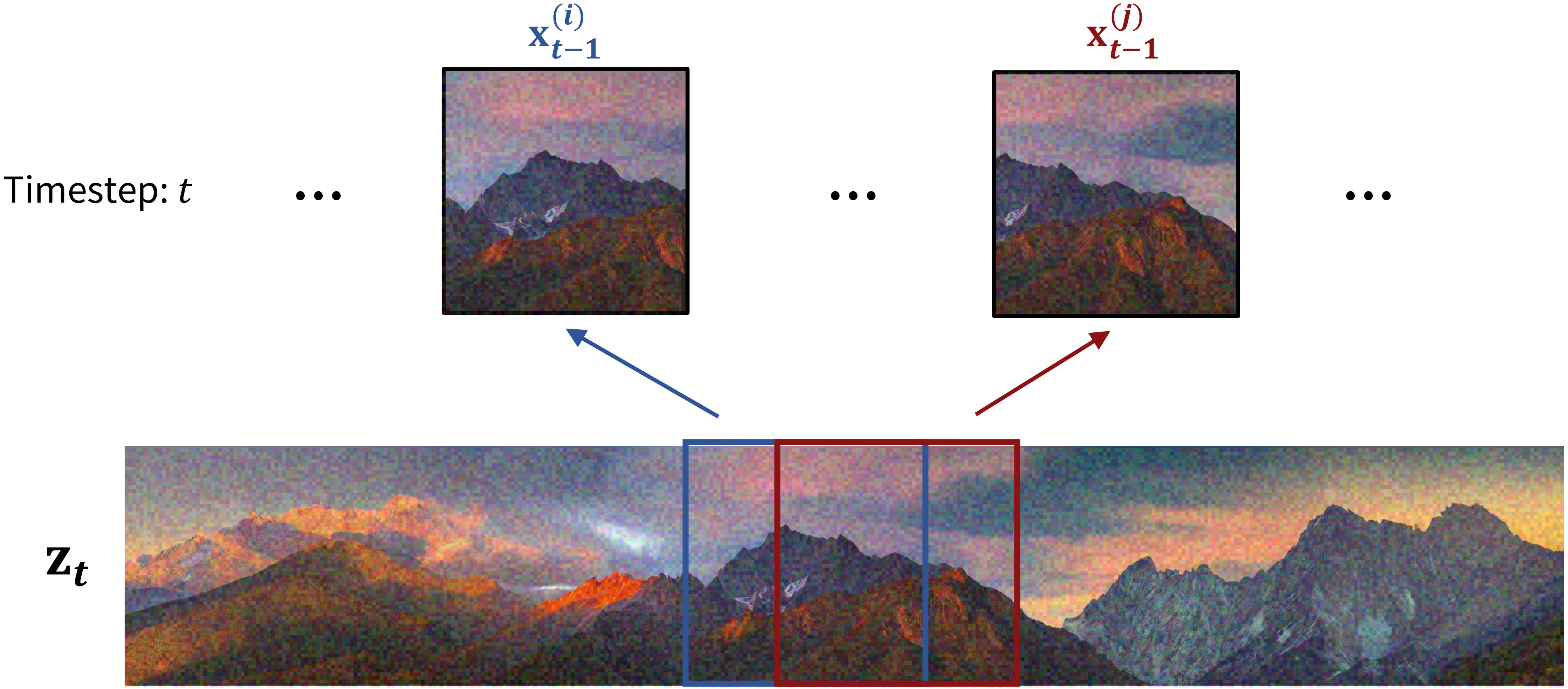


"A photo of a mountain range at twilight"

Bar-Tal et al., MultiDiffusion: Fusing Diffusion Paths for Controlled Image Generation, ICML 2023.

Joint Diffusion [MultiDiffusion, Bar-Tal *et al.*]

Crop the full latent to obtain the latent for each window.

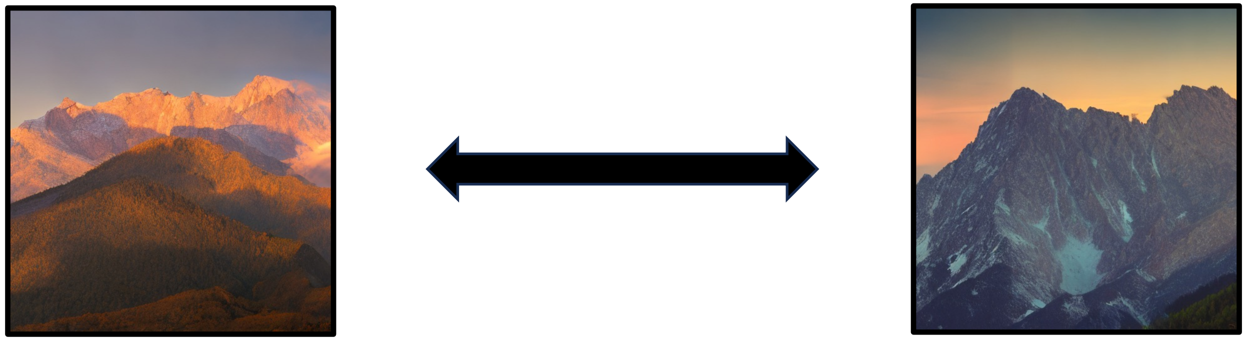


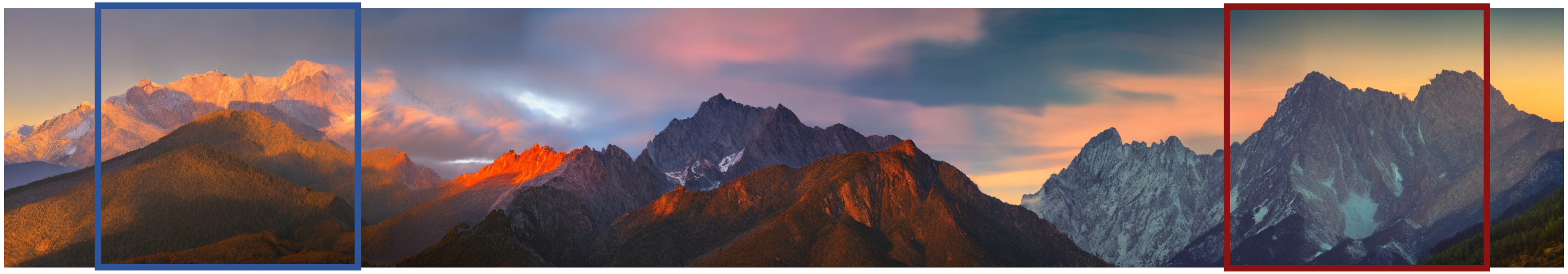
“A photo of a mountain range at twilight"

Bar-Tal et al., MultiDiffusion: Fusing Diffusion Paths for Controlled Image Generation, ICML 2023.

Joint Diffusion [MultiDiffusion, Bar-Tal *et al.*]

The final output is not coherent.



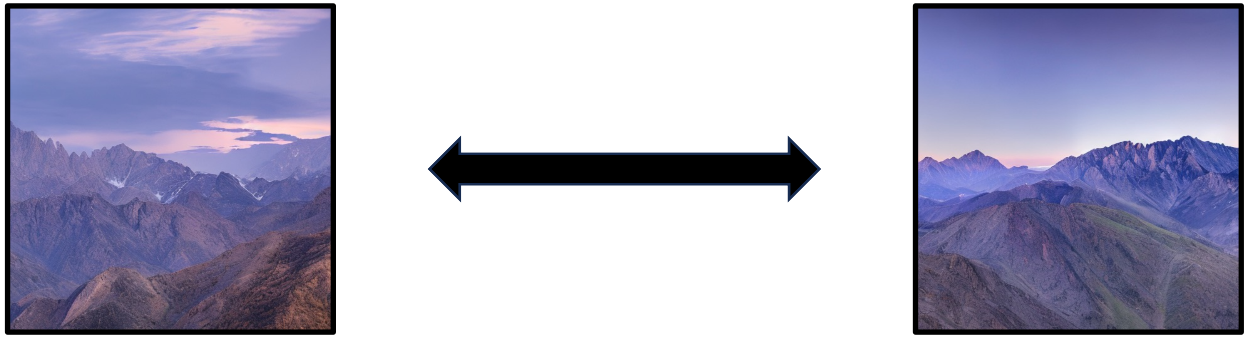


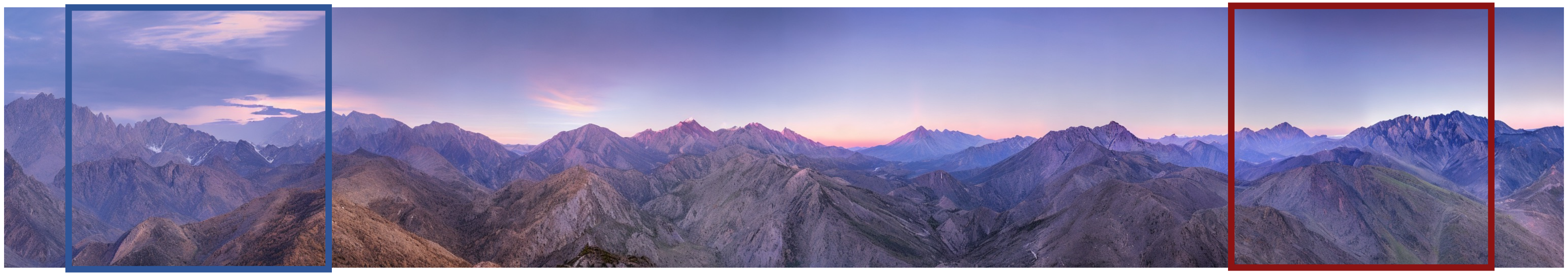
"A photo of a mountain range at twilight"

Bar-Tal et al., MultiDiffusion: Fusing Diffusion Paths for Controlled Image Generation, ICML 2023.

SyncDiffusion: Synchronized Joint Diffusions

Generate perceptually coherent images in arbitrary sizes.

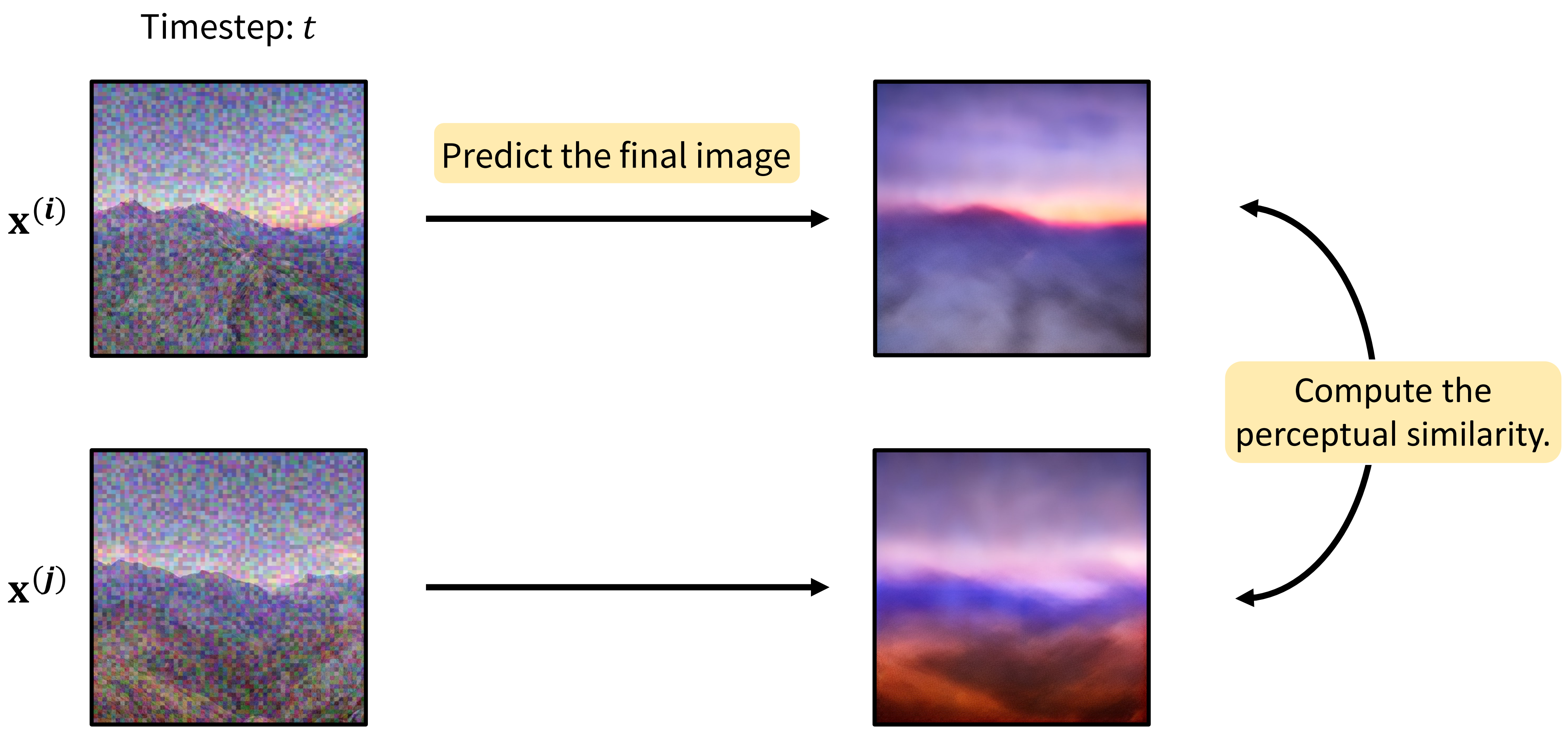




"A photo of a mountain range at twilight"

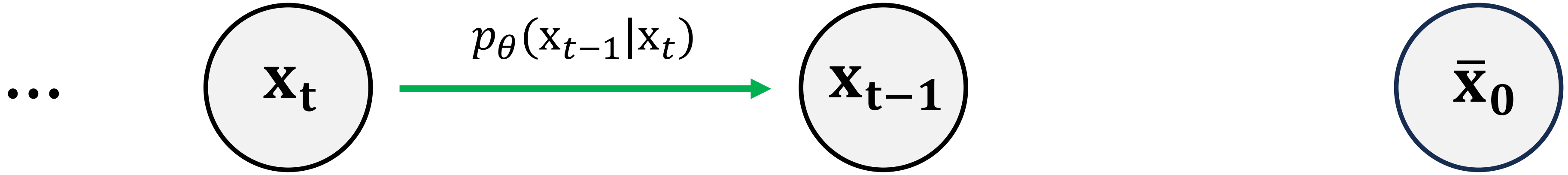
Key Idea

Compute the coherence in advance based on foreseen output images.



Background: DDIM [Denoising Diffusion Implicit Models]

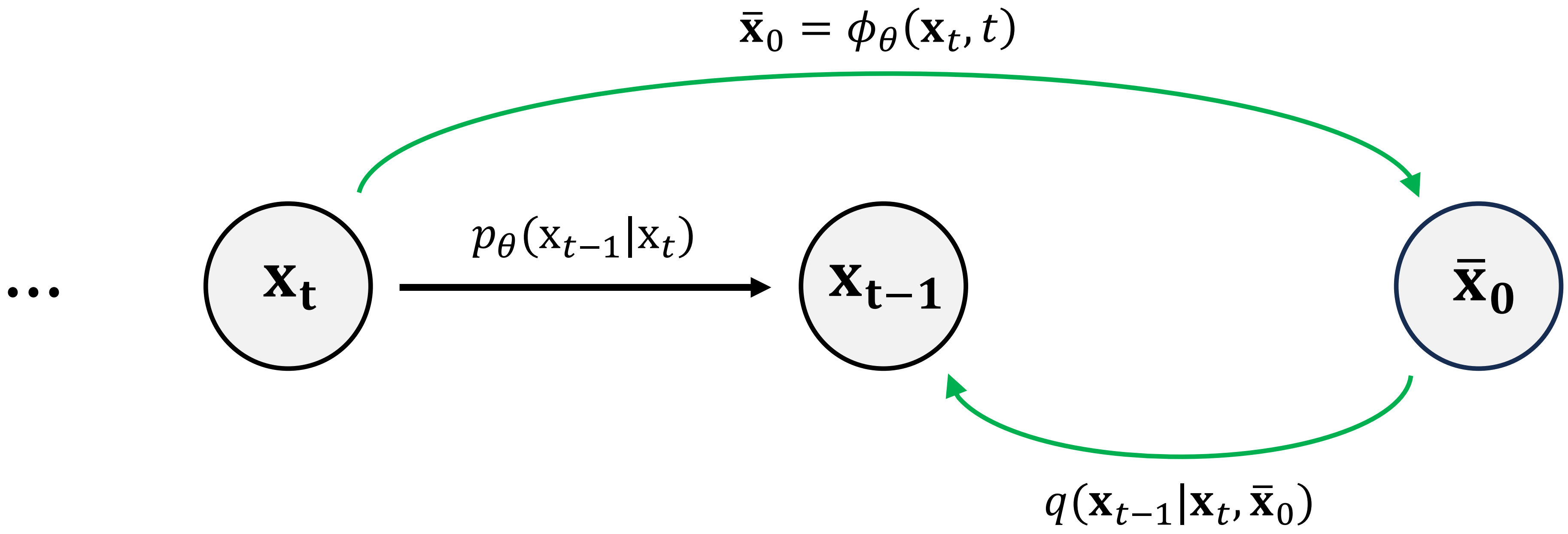
Transition from x_t to $\mathrm{x}_{\mathrm{t}-1}$ is conditioned on both $\mathbf{x}_{\mathbf{t}}$ and $\bar{\mathbf{x}}_{\mathbf{0}}$,
where $\bar{\mathbf{x}}_{\mathbf{0}}$ is the predicted denoised output given $\mathbf{x}_{\mathbf{t}}$ and timestep t.



Song et al., Denoising Diffusion Implicit Models, ICLR 2021.

Background: DDIM [Denoising Diffusion Implicit Models]

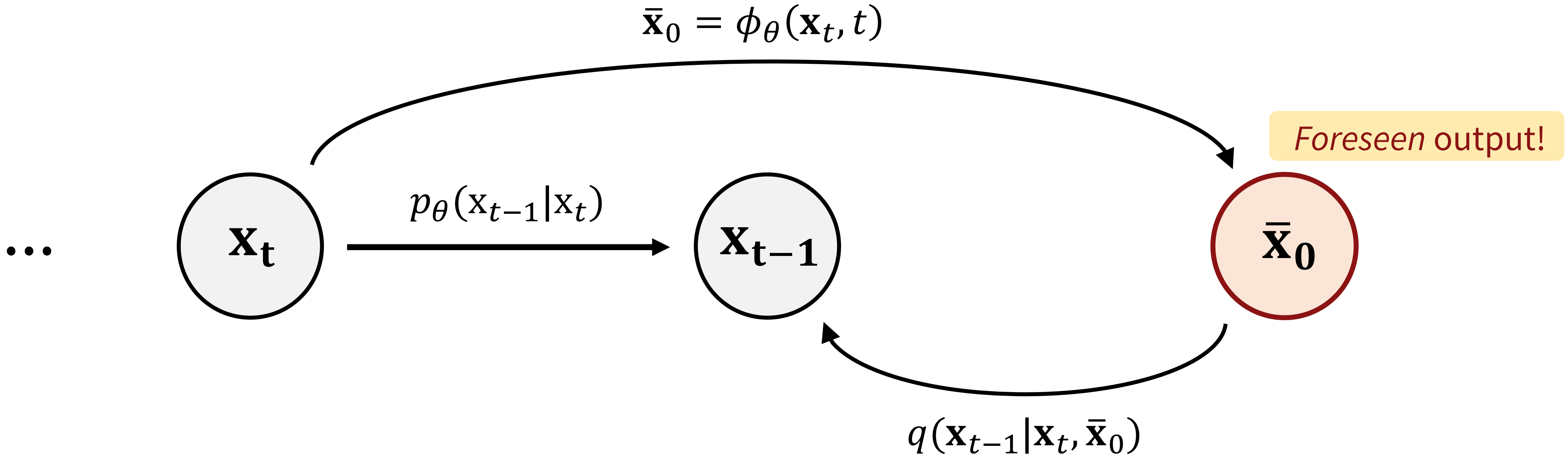
Transition from x_t to $\mathrm{x}_{\mathrm{t}-1}$ is conditioned on both $\mathbf{x}_{\mathbf{t}}$ and $\bar{\mathbf{x}}_{\mathbf{0}}$,
where $\bar{\mathbf{x}}_{\mathbf{0}}$ is the predicted denoised output given $\mathbf{x}_{\mathbf{t}}$ and timestep t.



Song et al., Denoising Diffusion Implicit Models, ICLR 2021.

Background: DDIM [Denoising Diffusion Implicit Models]

Transition from x_t to x_{t-1} is conditioned on both $\mathbf{x_t}$ and $\bar{\mathbf{x}}_\mathbf{0}$,
where $\bar{\mathbf{x}}_\mathbf{0}$ is the predicted denoised output given $\mathbf{x_t}$ and timestep t.



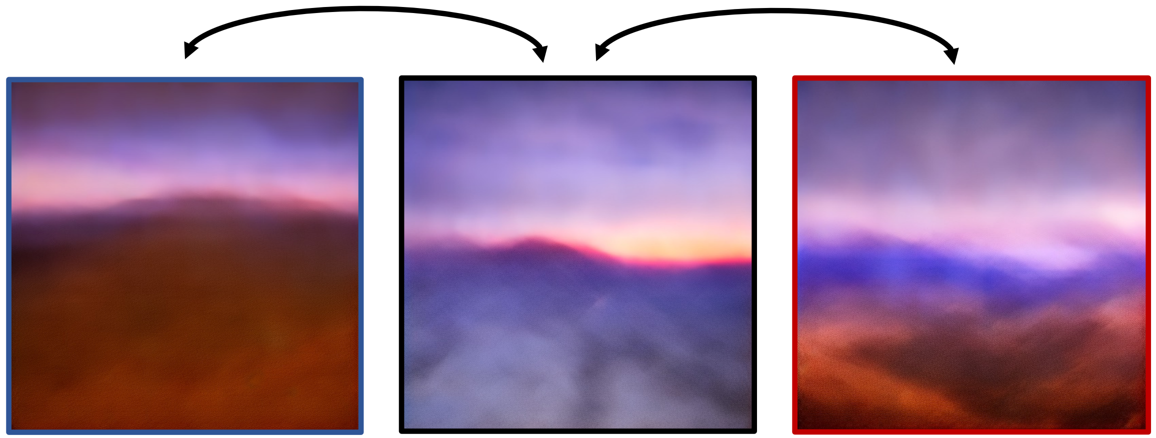
Song et al., Denoising Diffusion Implicit Models, ICLR 2021.

Observation

Perceptual similarity loss (i.e. LPIPS[1]) across foreseen images is aligned with that of the final images.

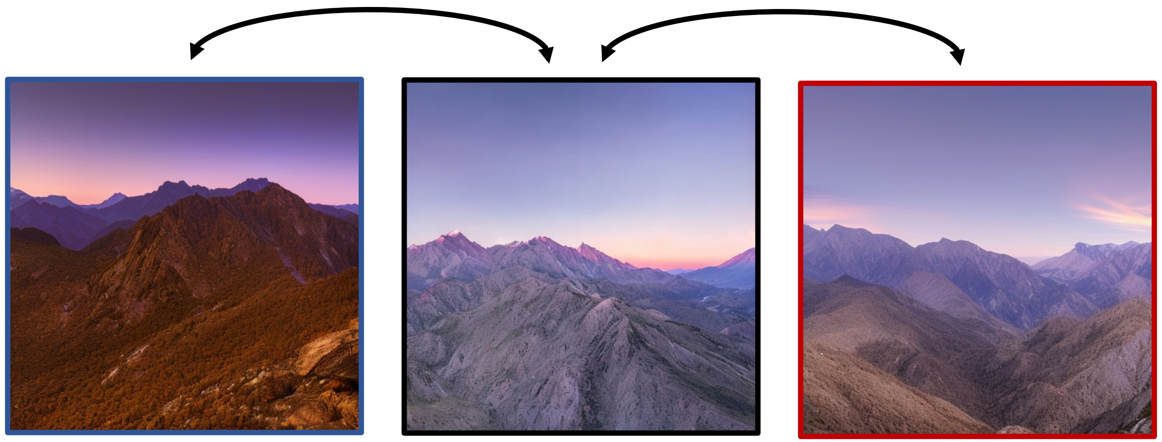
Foreseen outputs ($\bar{\mathbf{x}}_0$)

$L = 0.542 > L = 0.350$



Final outputs ($\mathbf{x}_0$)

$L = 0.591 > L = 0.370$

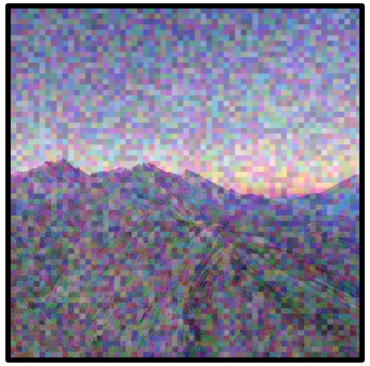


[1] Zhang et al., The Unreasonable Effectiveness of Deep Features as a Perceptual Metric, CVPR 2018

SyncDiffusion

Timestep: t

$\mathbf{x}_t^{(i)}$



$\mathbf{x}_t^{(0)}$

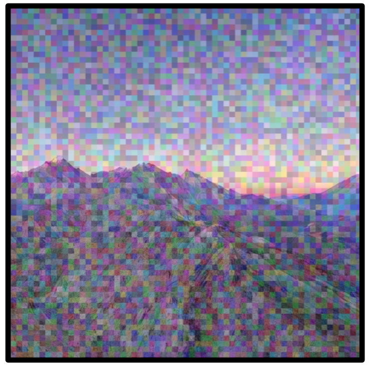
SyncDiffusion

(1) Predict foreseen output: $\bar{\mathbf{x}}_{\mathbf{0}}^{(i)} = \phi_\theta\left(\mathbf{x}_t^{(i)}, t\right)$

(2) Decode latent to image: $D\left(\bar{\mathbf{x}}_{\mathbf{0}}^{(i)}\right)$

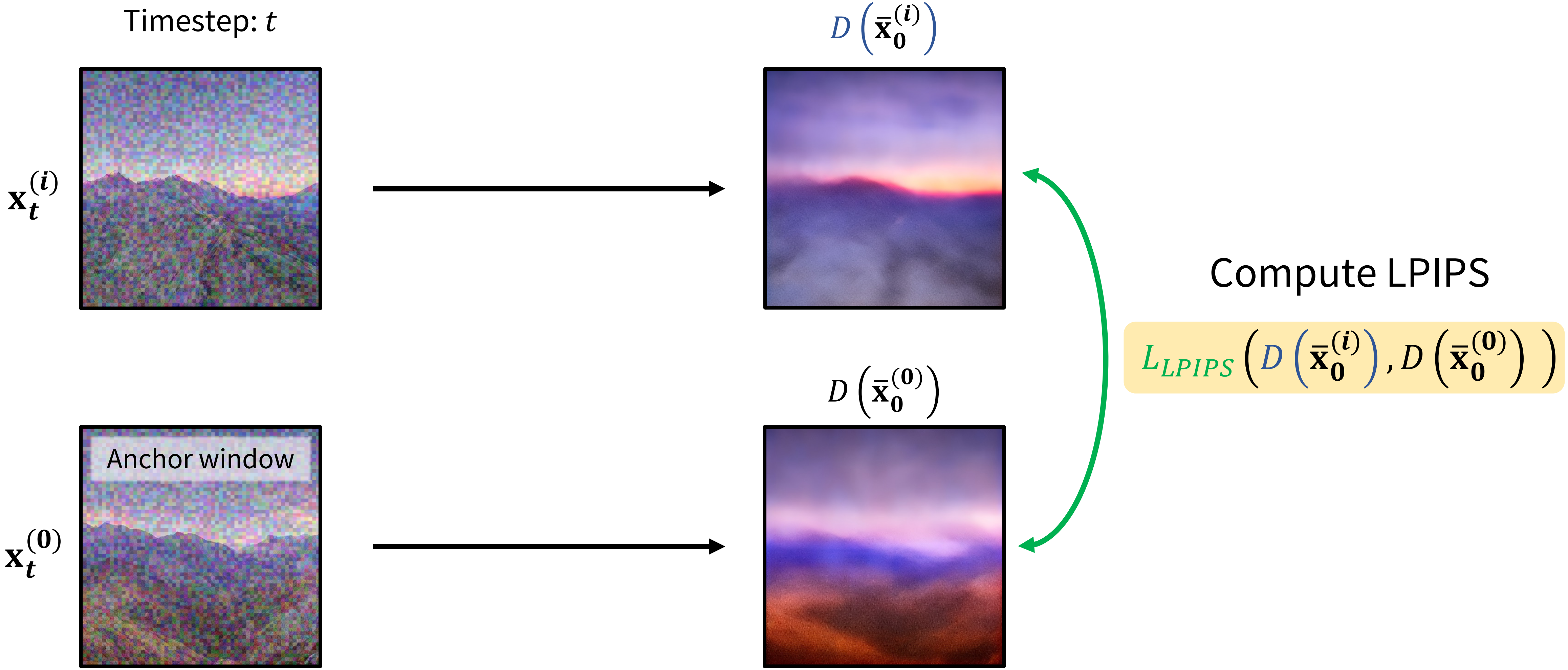
Timestep: t

$\mathbf{x}_t^{(i)}$

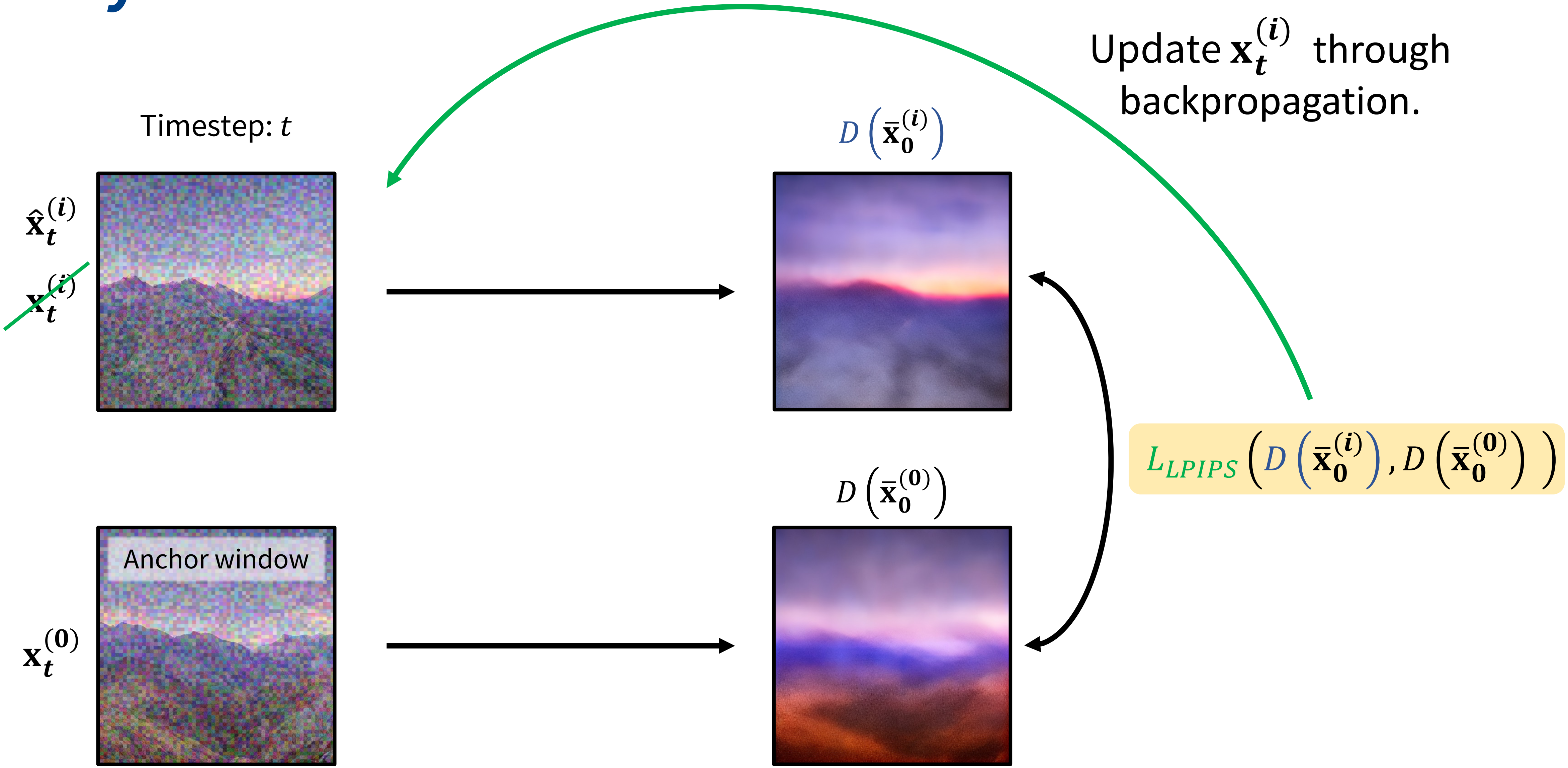


$\mathbf{x}_t^{(0)}$

SyncDiffusion



SyncDiffusion



Qualitative Results: Text-to-Panorama

MultiDiffusion (Bar-Tal et al.)

SyncDiffusion (Ours)

“Skyline of New York City”

Qualitative Results: Text-to-Panorama

MultiDiffusion (Bar-Tal et al.)

SyncDiffusion (Ours)

"A photo of a rock concert"

Qualitative Results: Text-to-Panorama

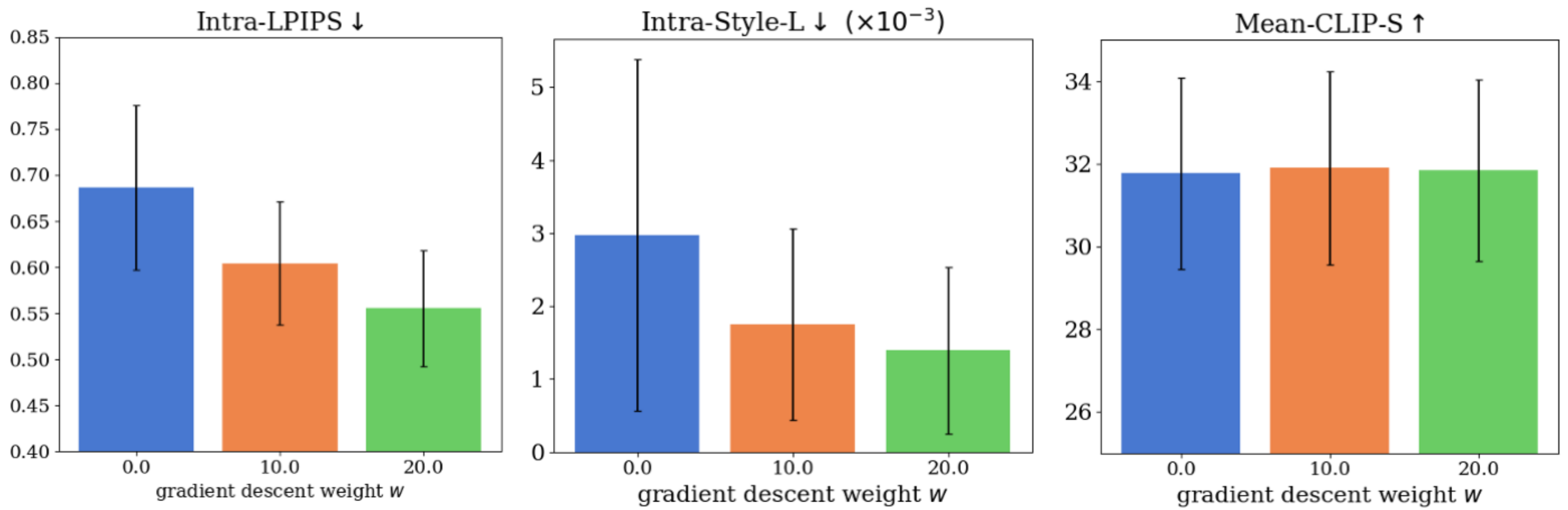
MultiDiffusion (Bar-Tal et al.)

SyncDiffusion (Ours)

“An illustration of a beach in La La Land style”

Quantitative Results

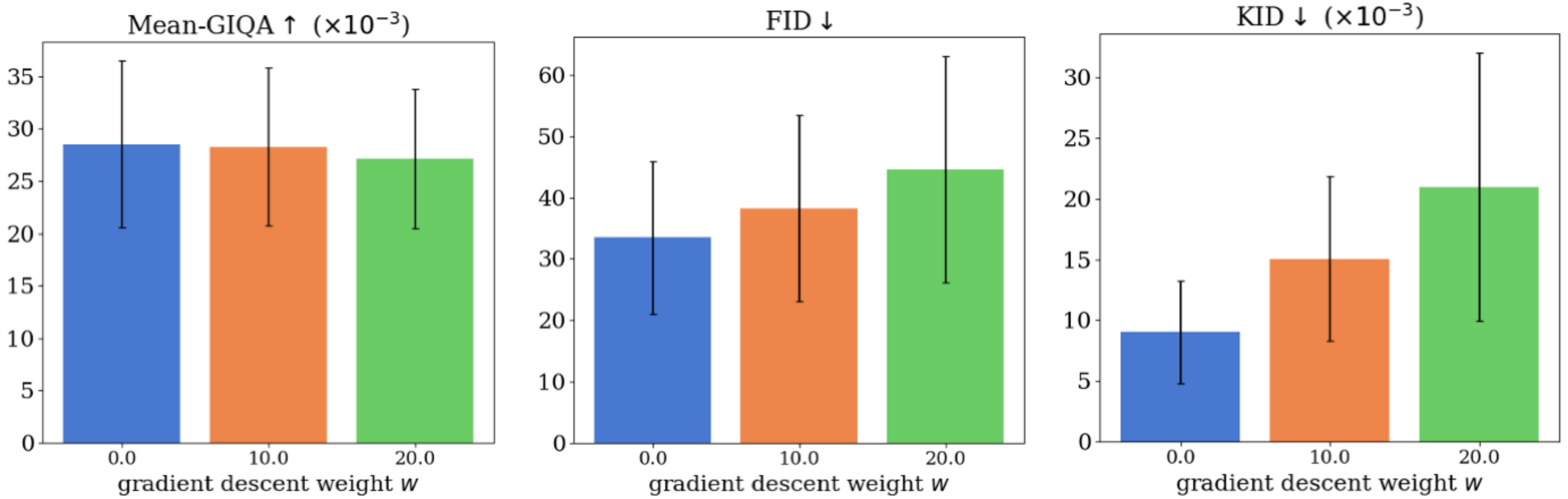
Coherence (LPIPS[1], Style Loss[2]) is improved while preserving the prompt compatibility (CLIP-S[3]) as the gradient descent weight w increases.



[1] Zhang et al., The Unreasonable Effectiveness of Deep Features as a Perceptual Metric, CVPR 2018.
[2] Gatys et al., Image style transfer using convolutional neural networks, CVPR 2016.
[3] Hessel et al., CLIPScore: A Reference-free Evaluation Metric for Image Captioning, EMNLP 2021.

Quantitative Results

Fidelity (GIQA[1]) is preserved, while diversity (FID[2],KID[3]) is slightly compromised as the gradient descent weight w increases.



[1] Gu et al., GIQA: Generated Image Quality Assessment, ECCV 2020.
[2] Heusel et al., GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium, NeurIPS 2018.
[3] Bińkowski et al., Demystifying MMD GANs, ICLR 2018.

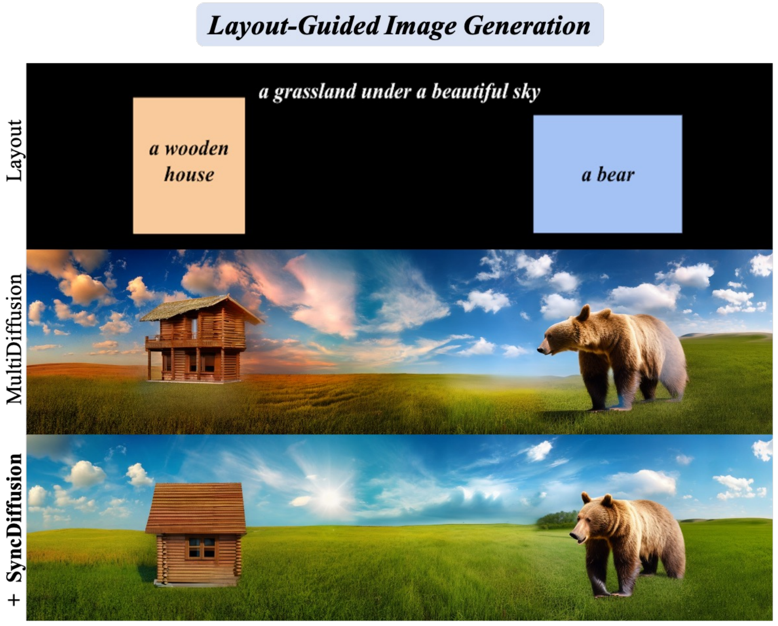
User Study

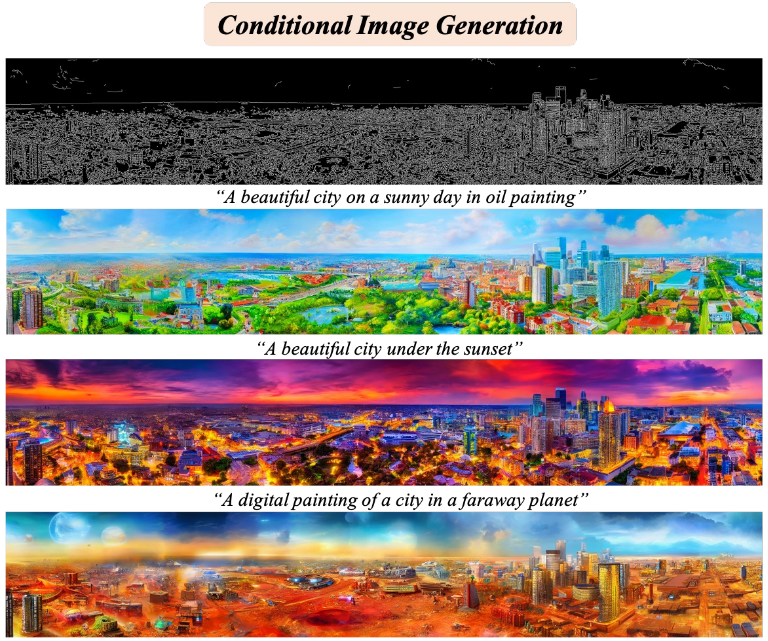
SyncDiffusion was preferred over the baseline for questions about coherence, image quality and prompt compatibility.

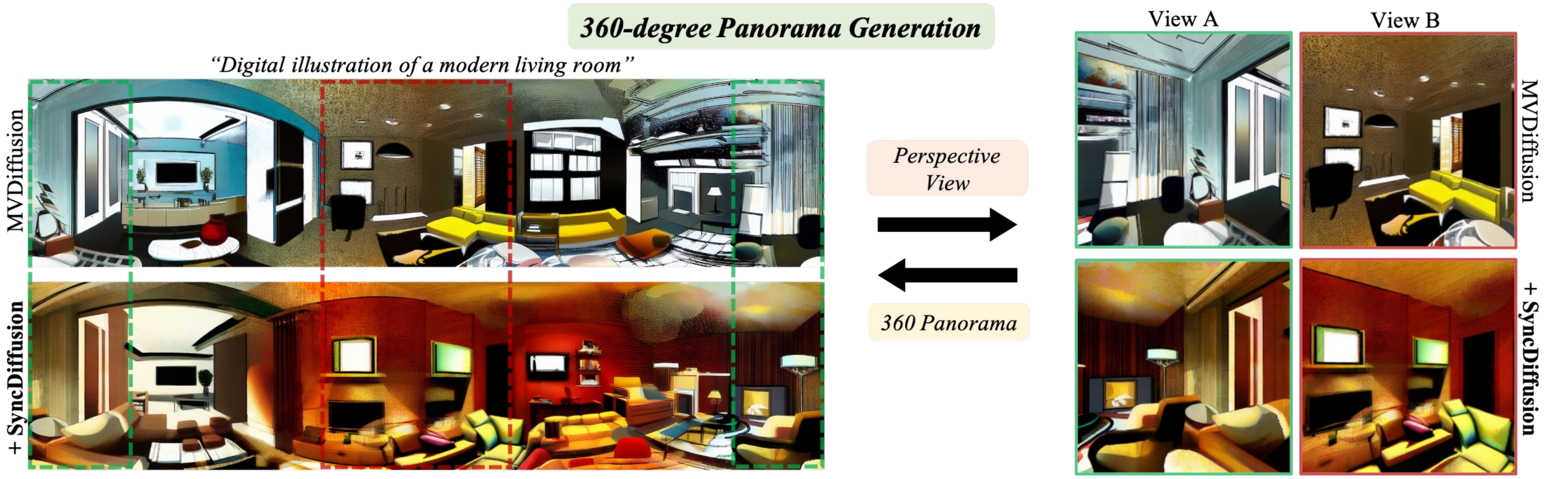
	Coherence (%)	Image Quality (%)	Prompt Compatibility (%)
MultiDiffusion[1]	33.65	42.81	40.50
SyncDiffusion (Ours)	**66.35**	**57.19**	**59.50**

[1] Bar-Tal et al., MultiDiffusion: Fusing Diffusion Paths for Controlled Image Generation, ICML 2023.

Plug-and-Play Applications







Bar-Tal et al., MultiDiffusion, ICML 2023.
Zhang et al., ControlNet, ICCV 2023.
Tang et al., MVDiffusion, NeurIPS 2023.

NEURAL INFORMATION PROCESSING SYSTEMS

SyncDiffusion: Coherent Montage via Synchronized Joint Diffusions

Session **3** | Poster #**532**

Project Page: https://syncdiffusion.github.io/

NeurIPS 2023

Yuseung Lee   Kunho Kim   Hyunjin Kim   Minhyuk Sung