MERGING AND SPLITTING DIFFUSION PATHS FOR SEMANTICALLY COHERENT PANORAMAS

Fabio Quattrini, Vittorio Pippi, Silvia Cascianelli, Rita Cucchiara
{fabio.quattrini, vittorio.pippi, silvia.cascianelli, rita.cucchiara}@unimore.it

Try the code!

Try the demo!

Overview

Diffusion models are the State-of-the-Art for text-to-image generation, and increasing research effort has been dedicated to adapting the inference process of pretrained diffusion models to achieve ***zero-shot*** capabilities. An example is the generation of long images, which has been tackled in recent works by combining strided diffusions over overlapping latent features. This yield perceptually aligned but semantically incoherent panoramas. ***We propose the Merge-Attend-Diffuse operator*** (MAD), pluggable into different types of diffusion models featuring attention operations, and an inference-time strategy for generating ***perceptually and semantically*** **coherent** panoramas.

MAD is modular: use it where and when you want

MAD can be applied in different blocks of the backbone and for a variable portion of the diffusion denoising chain, obtaining the desired tradeoff between uniformity and variability.

MAD applied in ***different backbone blocks:***

MAD applied for a ***different fraction of the denoising chain:***

MAD is effective and efficient

We perform ***quantitative comparison on standard panorama generation literature prompts*** and compare with direct inference across multiple resolutions. For MAD, we sweep over several application tresholds, quantifying the tradeoff between uniformity and variability.

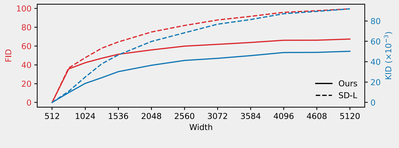
	mCLIP ↑	I-LPIPS ↓	FID ↓	KID ↓
SD [1]	31.63	0.74	28.31	<0.01
SD-L [1]	32.01	0.50	87.64	76.83
SD-L+Attn-S [5]	32.02	0.52	80.16	67.54
MD [2]	31.77	0.69	33.52	9.04
SyncD [3]	31.84	0.56	44.60	21.00
MAD (τ=0)	31.65	0.64	34.51	9.19
MAD (τ=5)	31.86	0.59	38.10	13.75
MAD (τ=15)	32.03	0.56	48.52	27.15
MAD (τ=25)	32.15	0.53	61.76	43.31
MAD (τ=40)	32.16	0.50	86.20	76.15
MAD (τ=50)	32.14	0.49	98.01	91.51

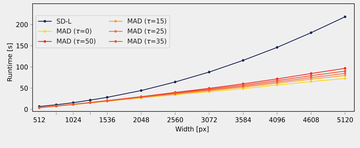
	mCLIP ↑	I-LPIPS ↓	FID ↓	KID ↓
2 Inference Steps				
LCD [4]	30.57	0.53	26.82	<0.01
LCD-L [4]	30.77	0.47	49.98	44.30
MAD (τ=0)	30.75	0.52	27.96	13.39
MAD (τ=1)	30.97	0.50	35.70	23.74
MAD (τ=2)	30.87	0.47	52.88	48.04
4 Inference Steps				
LCD [4]	31.37	0.55	29.06	<0.01
LCD-L [4]	31.30	0.50	55.52	51.72
MAD (τ=0)	31.36	0.55	31.35	13.28
MAD (τ=2)	31.48	0.52	38.78	27.42
MAD (τ=4)	31.41	0.48	62.82	61.70

We generate ***1000 complex-scene prompts*** (GPT1k) and perform a ***quantitative comparison*** with different baselines.

	mCLIP ↑	I-LPIPS ↓	FID ↓	KID ↓
SD [1]	32.45	0.67	49.32	<0.01
SD-L [1]	31.89	0.52	68.03	6.61
SD-L+Attn-S [5]	32.03	0.53	65.08	5.47
MD [2]	32.46	0.63	49.41	0.42
SyncD [3]	32.34	0.55	53.52	1.11
MAD	32.47	0.58	54.44	1.28
MAD+Attn-S [5]	32.40	0.57	55.58	1.82

Compared to direct inference methods ***MAD scales better*** by performing split convolutions for all timesteps and split attentions for the timesteps after the user-defined threshold.



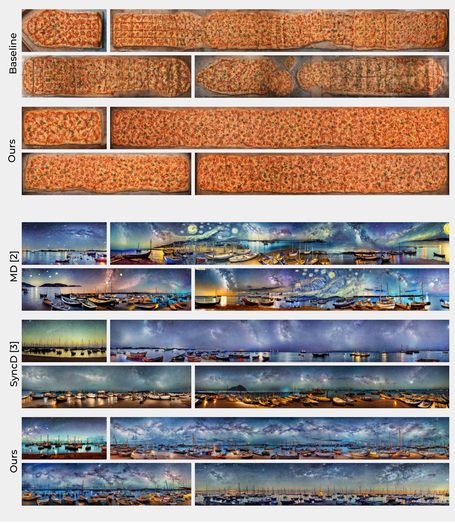


operator MAD $[\mathbf{v}^1_{t,j-1}, \dots, \mathbf{v}^I_{t,j-1}]$:
  $\mathbf{v}_{t,j-1} = \textsc{Merge}([\mathbf{v}^1_{t,j-1}, \dots, \mathbf{v}^I_{t,j-1}])$
  $\mathbf{v}_{t,j} = \textsc{Attention}_j(\mathbf{v}_{t,j-1})$
  $[\mathbf{v}^1_{t,j}, \dots, \mathbf{v}^I_{t,j}] = \textsc{Split}(\mathbf{v}_{t,j})$
  Result: $[\mathbf{v}^1_{t,j}, \dots, \mathbf{v}^I_{t,j}]$
end operator

$t = T$
$\mathbf{x}_t \sim \mathcal{N}(0, \mathbf{I})$
while $t > 0$ **do**
  $[\mathbf{v}^1_{t,0}, \dots, \mathbf{v}^I_{t,0}] = \textsc{Split}(\mathbf{x}_t)$
  for $j = 1$ **to** J **do**
    $[\mathbf{v}^1_{t,j}, \dots, \mathbf{v}^I_{t,j}] =$
      $\text{MAD}(\textsc{Conv}_j([\mathbf{v}^1_{t,j-1}, \dots, \mathbf{v}^I_{t,j-1}]))$
  end
  $\mathbf{x}_{t-1} = \textsc{Merge}(\textsc{Conv}_{j+1}([\mathbf{v}^1_{t,j}, \dots, \mathbf{v}^I_{t,j}]))$
  $t = t - 1$
end
$\hat{\mathbf{x}} = \mathcal{D}(\mathbf{x}_0)$
Result: $\hat{\mathbf{x}}$

MAD works just fine on different aspect ratios, on LDMs and LCMs

We ***generate horizontal panorama images at various aspect ratios***, from 512x1024 to 512x4096. Top: we apply MAD to an LCM [4] model and compare with a baseline. Bottom: we apply MAD to an LDM [1] and show images generated by competitors, namely MultiDiffusion [2] and SyncDiffusion [3].



We perform the same ***resolution sweep on vertical images*** by applying MAD to the LDM [1] with resolutions from 1024x512 to 4096x512 and showing images generated by competitors, namely MultiDiffusion [2] and SyncDiffusion [3].

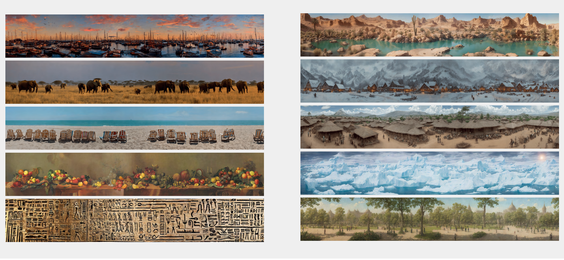
You can use MAD off-the-shelf for controlled image generation

We apply the MAD operator in ***other Plug&Play tasks for controllable image generation***, namely MultiDiffusion [2] and ControlNet [6], showing increased global coherence in the generated images

Not all prompts are panorama-friendly...

Inference-time panorama generation approaches struggle with scenes or objects that do not fit well with the specified aspect ratio and prompts where the base model itself does not produce good-quality results. We provide examples with the prompt *"A gothic cathedral nave"*, which does not fit well the horizontal aspect ratio. On a vertical canvas, results improve. Bottom: we provide examples with the prompt *"A fancy living room"*.

... But some are!



[1] Rombach, R., et al. "High-resolution image synthesis with latent diffusion models." CVPR 2022.
[2] Bar-Tal, O., et al. "Multidiffusion: Fusing diffusion paths for controlled image generation." ICML 2023.
[3] Lee, Y., et al. "Syncdiffusion: Coherent montage via synchronized joint diffusions." NeurIPS 2024.
[4] Luo, S., et al. "Latent consistency models: Synthesizing high-resolution images with few-step inference." arXiv 2023.
[5] Jin, Z., Shen, X., Li, B., Xue, X.: Training-free Diffusion Model Adaptation for Variable-Sized Text-to-Image Synthesis. NeurIPS 2023.
[6] Zhang, Lvmin, Anyi Rao, and Maneesh Agrawala. "Adding conditional control to text-to-image diffusion models." ICCV 2023.