DEMOFUSION-TURBO

Michele Minniti, Petr Sabel

Advanced Computer Vision Project 2024

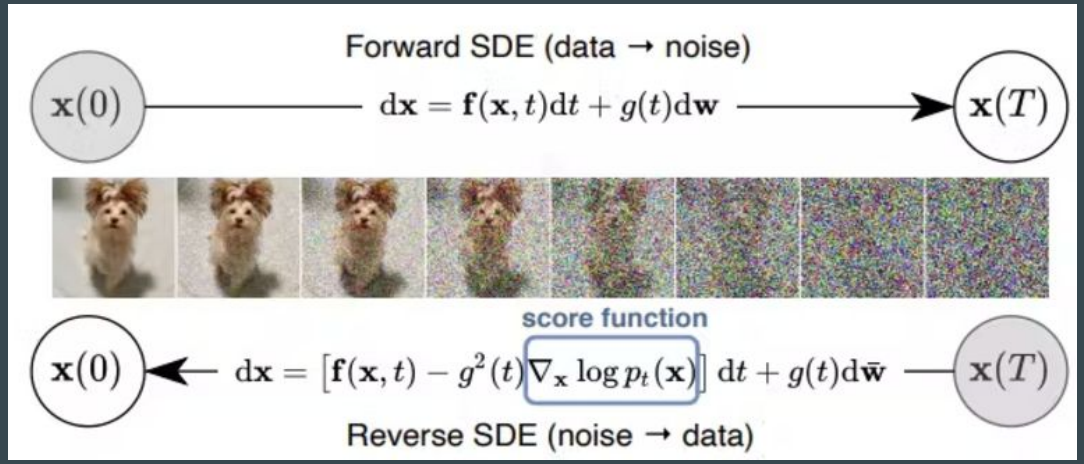
Introduction and Background

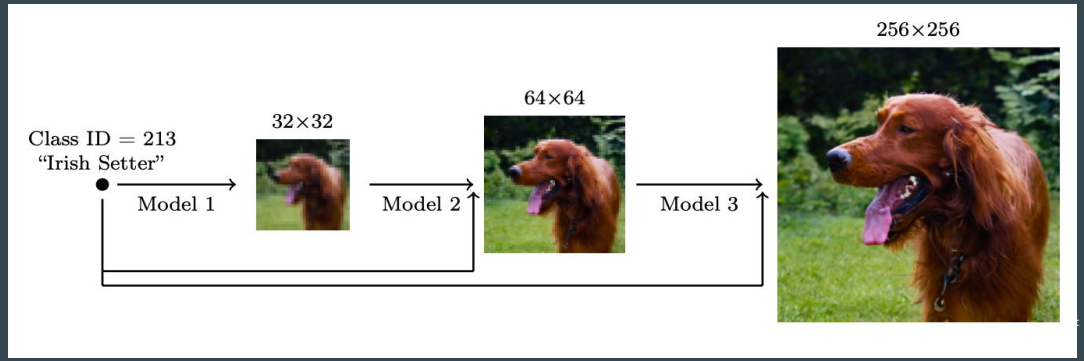
MOTIVATIONS

- Training Diffusion Models (DM) is extremely expensive
- High resolution image generation requires high investments in hardware
- DemoFusion is a method to generate high quality images on a (relatively) cheap hardware
- Based on Stable Diffusion XL (SDXL)
- High quality synthetic images from 1024x1024 to 4096x4096

Diffusion Models

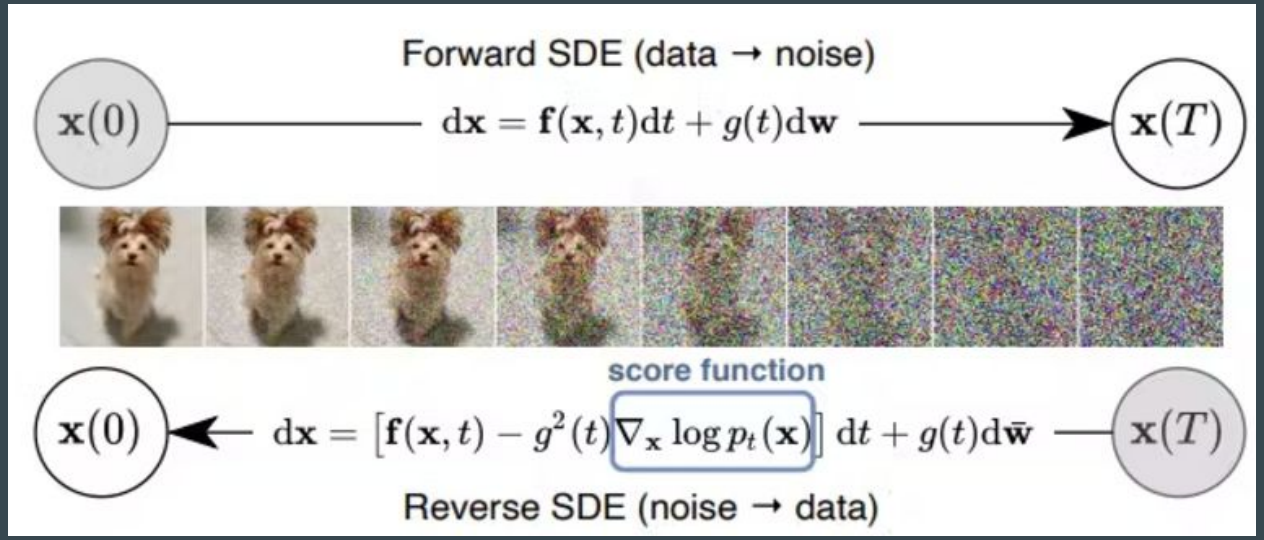
- DMs show great image quality and diversity in the generation.
- Using a pre-trained autoencoder Latent DM achieves great results with resolutions till 1024x1024.
- Cascaded DM can help us to increase final resolution.
    - A Cascaded DM comprises a pipeline of multiple DMs that increase the resolution of the image step by step.





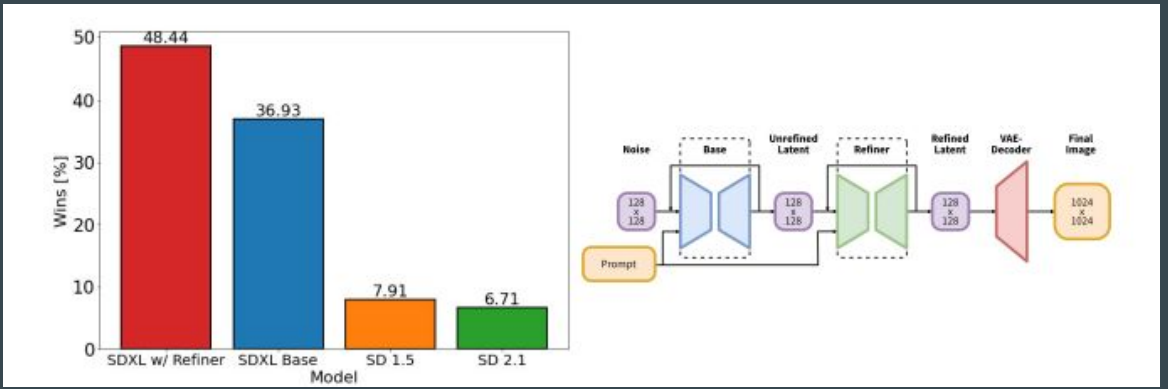
Diffusion Models

- DMs show great image quality and diversity in the generation.
- Using a pre-trained autoencoder Latent DM achieves great results with resolutions till 1024x1024.



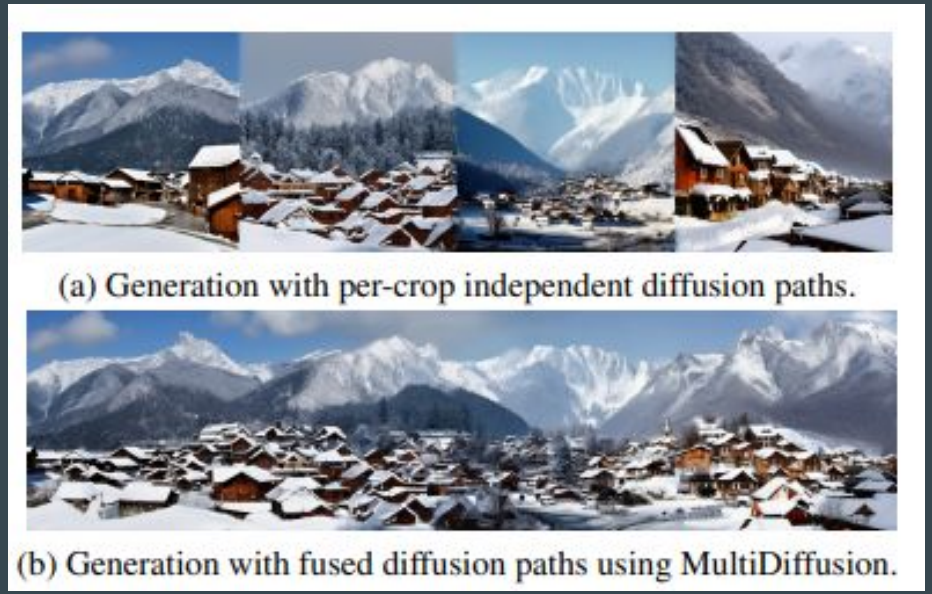
Stable Diffusion XL

- SDXL has consistently better performance than Stable Diffusion 1.5 and 2.1
- Two stage approach with an additional refinement model
  - First we generate initial latent of size 128x128
  - Afterwards we utilize a specialized high-resolution refinement model
  - SDXL and the refinement model use the same autoencoder
- Open source



Multidiffusion

- Framework that enables versatile and controllable image generation
- Pre-trained text-to-image DM
- The DM is applied to different regions of the image
- The regions have a shared set of parameters and constraints
- For each region we perform a denoising sampling step
- Then we make a global denoising sampling step reconciling all these different regions via a least squares optimal solution

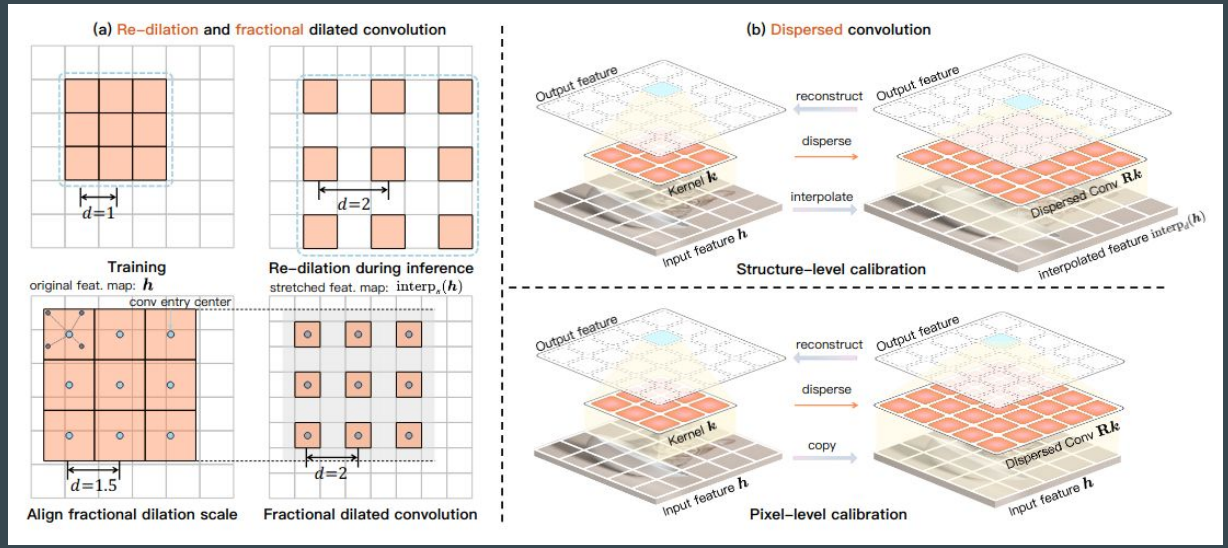


(a) Generation with per-crop independent diffusion paths.

(b) Generation with fused diffusion paths using MultiDiffusion.

SCALECRAFTER

- Concurrent model of DemoFusion
- The main problem is the absence of global semantic coherence in high resolution images generated by DMs due to the limited perception field of the convolutional kernels.
- They propose a re-dilation technique that can dynamically adjust the conv kernels.
- Convolution dispersion expands the receptive field by linearly transforming weights
- Noise-damped classifier-free guidance enhances details
- Object repetition is not totally absent



The solution of DemoFusion

Basic Methodology

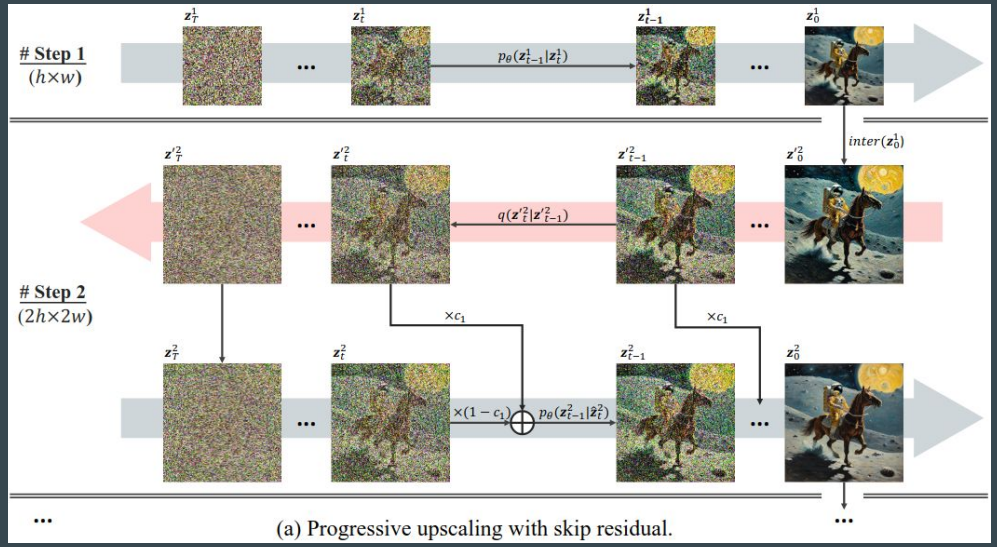
- An image x is encoded into a latent space $z = E(x)$ by a pre-trained autoencoder
- We gradually add noise over T time steps
- We then denoise to recover a cleaner version of z
- Multidiffusion extends this process to obtain higher resolution images:
  - We define a larger latent space
  - Then apply a shifted crop sampling and obtain a series of local latent representations
  - Then denoising is applied independently to these local z's
  - Eventually a higher resolution image is obtained decoding z_0 to x'
- This is not enough to obtain globally semantic coherent images since each patch is constrained only by the text condition and lacks awareness of the global context of the other patches
- In the next slides we'll discuss the solution found to this problem

Progressive upscaling

The goal is to generate images from low to high resolution, first creating a semantically coherent structure, and then enhancing the details of the image

- We break the generation phase in *S* phases
- First step is an “initialise-denoise”
- Next steps are “upsample-diffuse-denoise”
  - Upsample to the next resolution level
  - Introduce noise at the new scale
  - Denoise to obtain a clean image with new details

In this way we can compensate for the artificial interpolation based upsampling and gradually fill in more and more local details.

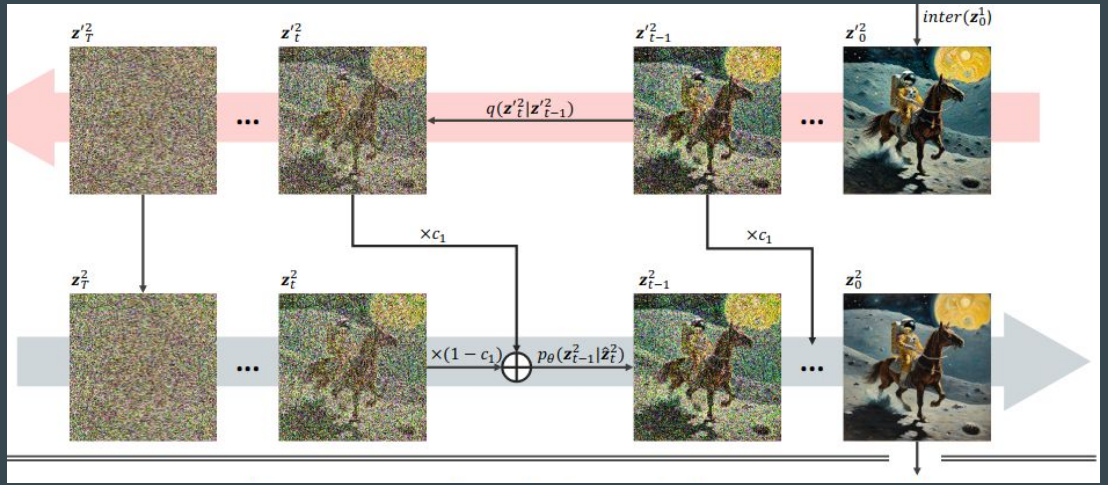


(a) Progressive upscaling with skip residual.

Skip residual

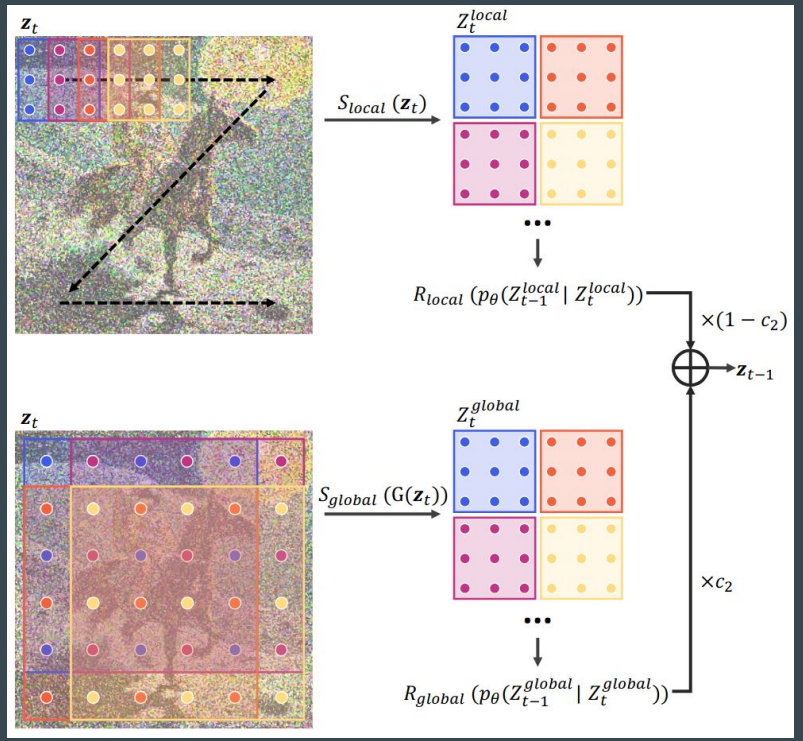
The goal is to obtain the better possible trade off between the preservation of the global structure of the image and the refinement of the details in the new resolution

- Skip residual is basically a weighted fusion of multiple “upsample-diffuse-denoise” loops with a series of different intersection timesteps *t*
- Combines the residual *z’* with the current latent representation *z*
- Early steps emphasize global structure (stronger noise residual influence)
- Later steps focus on refining local details (weaker noise residual influence)



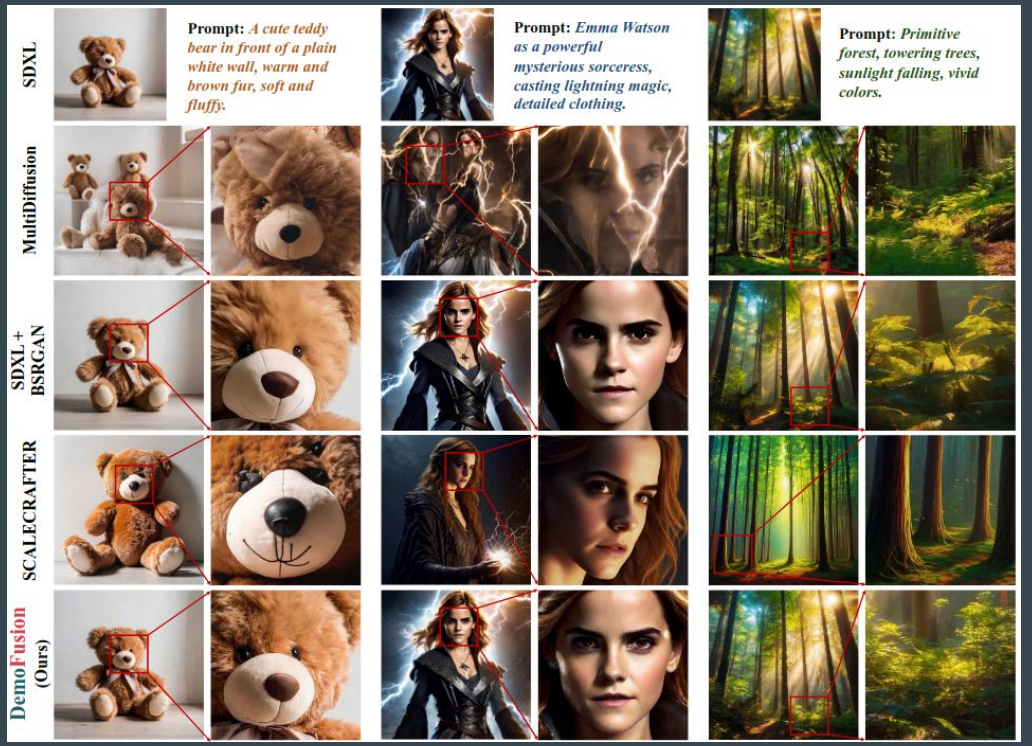
Dilated sampling

- Similar to SCALECRAFTER convolution dispersion
- But instead of dilating the kernel, they dilate the sampling within the latent representation
- Then the global denoising paths are processed the same way as local denoising path in MultiDiffusion
    - However there is a difference: global denoising paths don't present overlaps!
    - This leads to grainy images
    - => Apply a Gaussian filter to the latent representation!



Qualitative results

- MultiDiffusion tends to generate repetitive content lacking semantic coherence.
- SDXL+BSRGAN needs more detail for true high-resolution visuals beyond simple smoothing.
    - High-resolution generation cannot be substituted by simple image super-resolution.
- SCALECRAFTER partially addresses the issue of repetitive content, but directly dilating the convolutional kernels results in an overall image quality degradation, and local details exhibit many repetitive patterns.
- DemoFusion achieves both rich local details and strong global semantic coherence.



Quantitative evaluation

- **Inception score:** Measure image quality and diversity
  - Pretrained inception model predicts class probabilities for generated images
  - Better images => clear predictions $p(y|x)$
  - Diverse images => balanced marginal distribution $p(y)$
- **FID score:** Measure similarity between generated images and real world images
  - Fréchet inception distance quantify how close real and synthetic images are
  - Compares mean and covariance of real and synthetic features
- **CLIP score:** Evaluate semantic similarity between images and their textual description
  - Use CLIP model to compute similarity
  - Higher the score means better adherence to the prompt

Results for 4096x4096 images

Method	FID	IS	FID-crop	IS-crop	CLIP	Time (min)
SDXL	105.65	14.01	98.59	19.47	25.64	8
MultiDiffusion	97.98	13.84	79.45	19.73	28.62	15
SDXL + BSRGAN	**66.44**	**16.21**	77.20	22.42	**29.63**	1
SCALECRAFTER	87.50	15.20	84.36	20.32	29.04	19
DemoFusion	74.11	16.11	**70.34**	**24.28**	29.57	25

Experiments runned by the authors of the paper (on LAION-5B dataset)

Limitations

- The MultiDiffusion style inference method requires high computational load due to the overlapped denoise and progressive upsampling
- Depends on LDM performance and its previous knowledge
- Repetitive content is always a problem in the background of some images

Opportunities

- DemoFusion fuses multiple denoising paths, so we can implement each denoising step in mini batches preventing the exponential use of memory.
- The priors of current LDMs regarding image crops derive only from the general training scheme; training a custom LDM for a DemoFusion-like framework could be interesting.
- Progressive upscaling gives us the chance to use low resolution and intermediate images as previous of the final work.

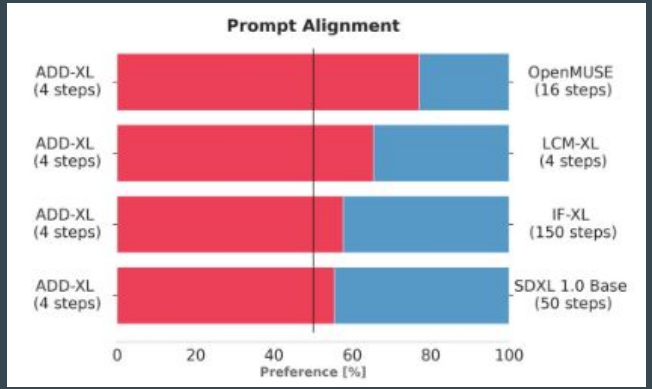
Our idea

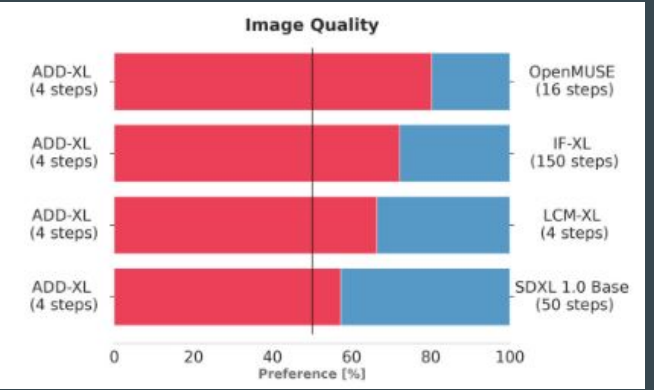
What we thought

- The main problem of DemoFusion respect to his competitors is definitely inference time
- Our idea is try to diminish it without a significant drop in quality
- How to obtain it?
  - **Change SDXL skeleton to a new pretrained model**
    - Many other models had many problem of adapting to the DemoFusion pipeline
    - SDXL-Turbo is distilled version of SDXL
    - SDXL-Turbo had overall the best results maintaining quality and leaning to faster results
  - **Quantization on the pre-trained autoencoder**
    - We quickly realized that the inference time became slower
    - The reason is that DemoFusion pipeline already applies multiple quantization and dequantization of the autoencoder, so it only increases the GPU computational load
    - Quantization on other components didn't seem profitable

SDXL-Turbo

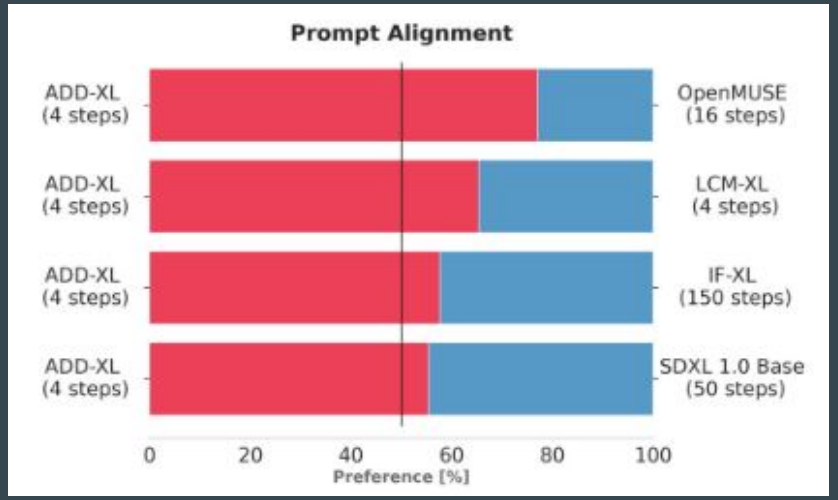
- Capable of outperforming other generators with only 1-4 sampling steps on low resolutions (512x512)
- Suitable for real time applications, faster inference time even with similar quantity of steps
- Capable of generating high quality images even at 1024x1024 as SDXL
- Adversarial Diffusion Distillation (ADD):
  - It composes methods from GANs and DMs
  - Adversarial Loss: A discriminator is trained to distinguish between real and generated images
  - Score Distillation Loss: Uses a teacher model as reference, the student approximates teacher's output
  - The total loss is a combination of the two losses

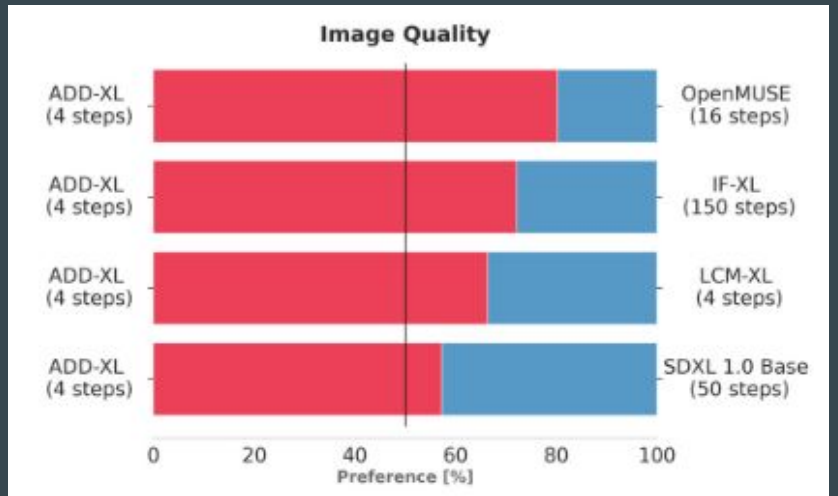




SDXL-Turbo

- Capable of outperforming other generators with only 1-4 sampling steps on low resolutions (512x512)
- Suitable for real time applications, faster inference time even with similar quantity of steps
- Capable of generating high quality images even at 1024x1024 as SDXL





Our Contribution

Our experiments

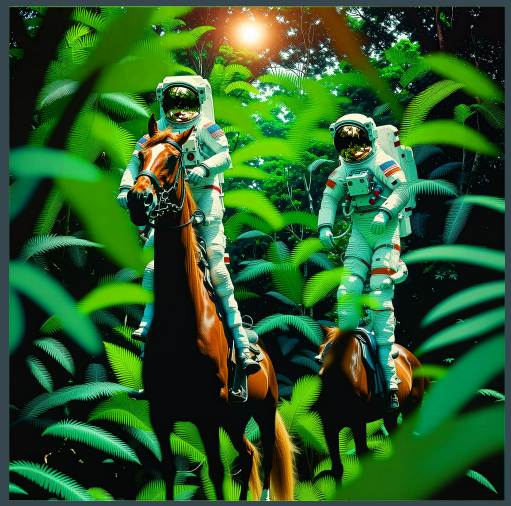
- The paper code already includes the version for systems with lower VRAM resources (around 8 GB).
- SDXL-Turbo: faster inference but trained to generate 512x512 images
  - Doing experiments we notice that the inference time for the same resolution is faster (with same amount of steps)
- Change the number of inference steps:
  - DemoFusion techniques are less effective with less steps
  - Turbo model works well with few steps
- Conditioning on both text and lower-resolution image
  - Better quality in initial images => better quality after upsampling

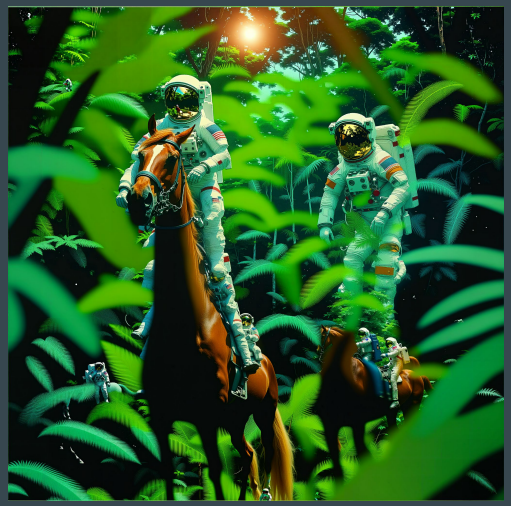
Our results

- After the first iteration the model enrichs the image details without creating repetitive content.
- The better starting image generates more accurate outputs.
- At lower resolution the results are outstanding.
- Increasing the resolution creates problems of coherence.

Few steps

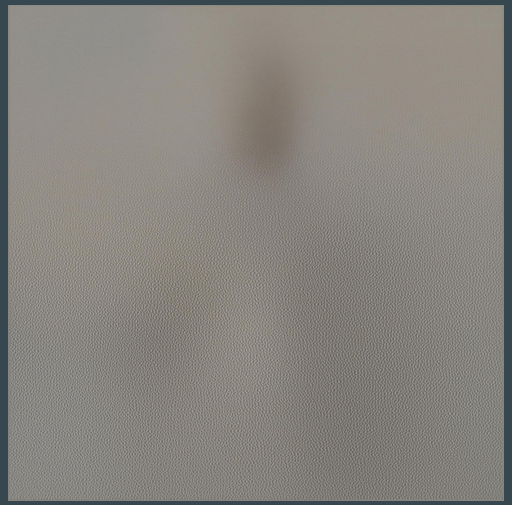
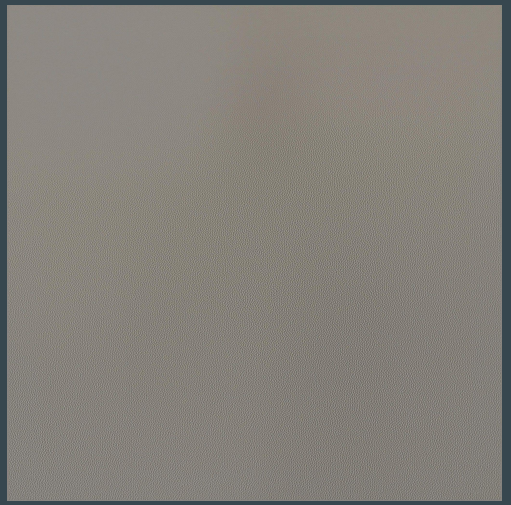
- Even with few steps the Turbo can add details but lacks semantic coherence
- The content depends a lot on the starting image





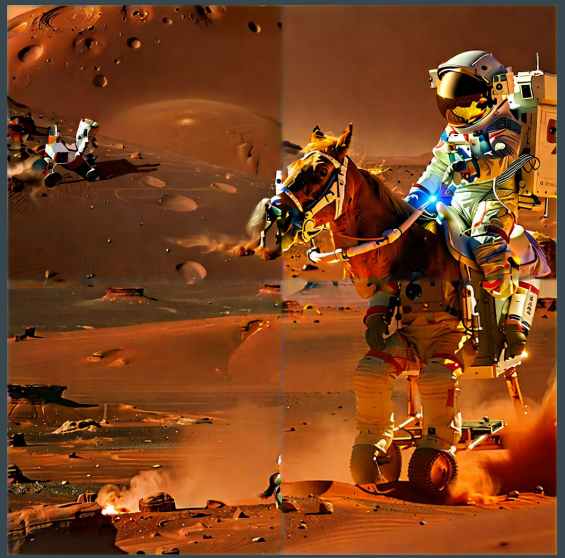
Few steps SDXL

- With few steps SDXL is slower and produces only noise
- As the number of steps increases the quality of SDXL increases dramatically while SDXL Turbo still presents problems of semantic coherence

Absence of global semantics

- With bad configuration parts of the image become too independent
  - Especially with small batch size and large stride



- Less steps produce more smooth samples and even small perturbation may produce repetitive content
- The repetitions tends to appear at same spots
- It would be interesting to study the nature of noise that produces these repetitions

Time results

- As the number of steps grows the advantages of SDXL Turbo become less relevant

Model	#Steps	Stride	Guidance Scale	Batch size	Resolution^2	Time (min)
SDXL	50	64	7.5	16	3072	33
SDXL	10	48	2.5	12	3072	18
SDXL Turbo	10	48	2.5	12	3072	12
SDXL Turbo	20	48	2.5	12	3072	20
SDXL Turbo	50	64	7.5	16	3072	31

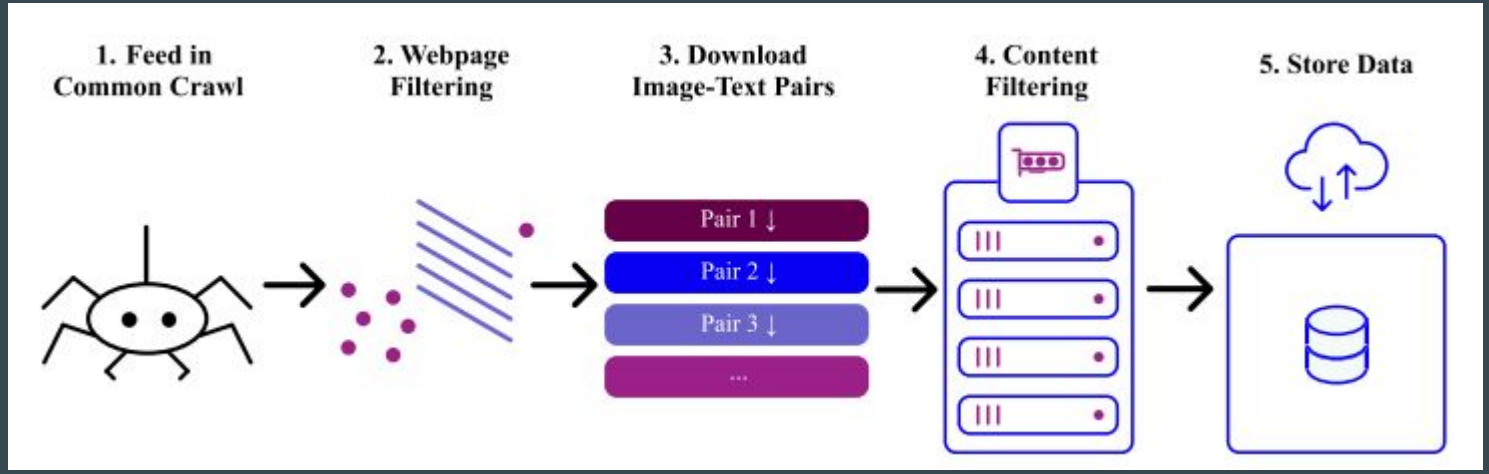
Experiments runned on Google Colaboratory

References

1. Du, R., Chang, D., Hospedales, T., Song, Y.-Z., and Ma, Z., “DemoFusion: Democratising High-Resolution Image Generation With No $$$”, Art. no. arXiv:2311.16973, 2023. doi:10.48550/arXiv.2311.16973.
2. Bar-Tal, O., Yariv, L., Lipman, Y., and Dekel, T., “MultiDiffusion: Fusing Diffusion Paths for Controlled Image Generation”, Art. no. arXiv:2302.08113, 2023. doi:10.48550/arXiv.2302.08113.
3. Dhariwal, P. and Nichol, A., “Diffusion Models Beat GANs on Image Synthesis”, Art. no. arXiv:2105.05233, 2021. doi:10.48550/arXiv.2105.05233.
4. Podell, D., “SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis”, Art. no. arXiv:2307.01952, 2023. doi:10.48550/arXiv.2307.01952.
5. Ho, J., Saharia, C., Chan, W., Fleet, D. J., Norouzi, M., and Salimans, T., “Cascaded Diffusion Models for High Fidelity Image Generation”, Art. no. arXiv:2106.15282, 2021. doi:10.48550/arXiv.2106.15282.
6. He, Y., “ScaleCrafter: Tuning-free Higher-Resolution Visual Generation with Diffusion Models”, Art. no. arXiv:2310.07702, 2023. doi:10.48550/arXiv.2310.07702.
7. Schuhmann, C., “LAION-5B: An open large-scale dataset for training next generation image-text models”, Art. no. arXiv:2210.08402, 2022. doi:10.48550/arXiv.2210.08402.
8. Barratt, S. and Sharma, R., “A Note on the Inception Score”, Art. no. arXiv:1801.01973, 2018. doi:10.48550/arXiv.1801.01973.
9. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S., “GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium”, Art. no. arXiv:1706.08500, 2017. doi:10.48550/arXiv.1706.08500.
10. Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., and Choi, Y., “CLIPScore: A Reference-free Evaluation Metric for Image Captioning”, Art. no. arXiv:2104.08718, 2021. doi:10.48550/arXiv.2104.08718.
11. Sauer, A., Lorenz, D., Blattmann, A., & Rombach, R. (2025). Adversarial diffusion distillation. In *European Conference on Computer Vision* (pp. 87-103). Springer, Cham.

QUESTIONS?

Dataset: LAION-5B



- 5.85 billion of CLIP-filtered image-text pairs
  - Primarily English but also contains non-identified languages
- Can be used to train CLIP-like models
- Often used to train generative models

- May overlap with test sets: the commonly used datasets can be part of this one, so performance may be unreasonably high
- Reference tag is not necessarily a good description of the image
- CLIP was used to filter image-text pairs, so the quality depends a lot on its performance.