Semantic Draw : Towards Real-Time Interactive Content Creation from Image Diffusion Models

Jaerin Lee, Daniel Sungho Jung, Kanggeon Lee, Kyoung Mu Lee

ECE & ASRI & IPAI, Seoul National University

CVPR Nashville JUNE 11-15, 2025

ComputerVisionLab Seoul National University

INTRODUCTION

TL;DR

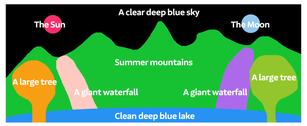
By marrying previously incompatible region-based diffusion & fast schedulers, Semantic Draw turns any image diffusion model into a live paintbrush tool that let you brush semantic intent—not just color pixels.

Motivation

(1) MultiDiffusion (51m 39s)

(2) MD + LCM Acceleration (4m 47s)

(3) **Ours** (59s)



(4) Multi-region prompt input

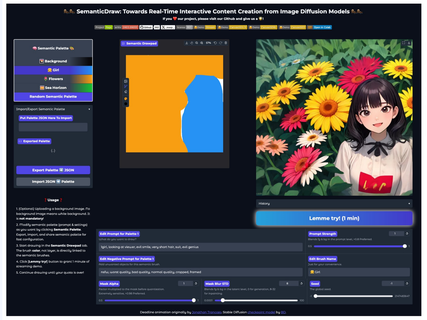
- ❑ Current region-based image diffusion pipelines do not scale to large canvas with many regional prompts.
- ❑ Naïvely adding acceleration modules (LCM, SDXL-Lightning, Flash-SD3, Hyper-SD etc.) does not work because they are not developed in aware of each other.
- ❑ This limits the practical power of BOTH works: regional control/acceleration.

Main Contributions

- ✓ We establish compatibility between region-based image diffusion pipelines & accelerated schedulers, speeding up the generation up to x50 faster with higher stability ready for real-time applications.
- ✓ We propose multi-prompt stream batch architecture to maximize the throughput & hide the latency of generation.
- ✓ Based on this, we propose a new drawing paradigm, "semantic drawing," for any image diffusion pipelines.

PROPOSED METHOD

Semantic Draw (Streaming)

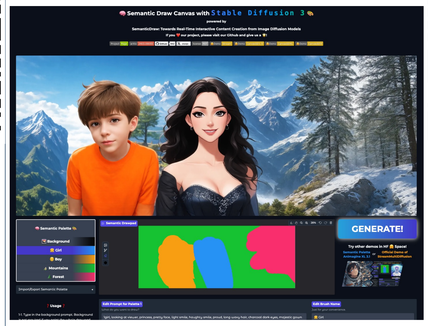


Try this out!

Instruction

1. In Semantic Palette, Edit brush prompts to choose what to draw.
2. Click [Lemme Try!] and get your 1-minute quota.
3. Start draw with brushes that paints meanings.

Semantic Draw (Large Canvas)



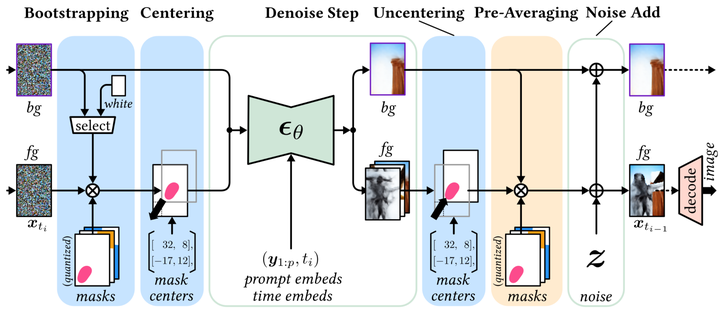
Take a look!

Instruction

1. In Semantic Palette, Edit brush prompts to choose what to draw.
2. In Drawpad, draw masks with semantic brushes.
3. Click [Generate] button to get your result.

UNDER THE HOOD

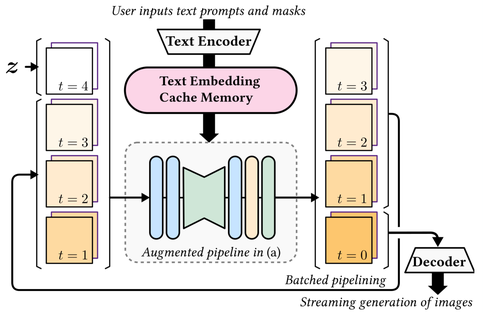
Redesigning the Pipeline



(a) Any image diffusion pipeline (green) is augmented with compatibility module.

- ❑ Bootstrap: In the early stages (1-3), the intermediate latent from each prompts is masked by its companion masks. Each masked latents are centered to exploit object-centeredness bias of diffusion models. After denoising, foregrounds are uncentered and merged.
- ❑ Scheduler Compatibility: Euler-type (only denoising) and Langevin-type (noise is added to each step) samplers are treated equally by delaying the noise addition step after merge.

Multi-Prompt Stream Batch



(b) Streaming by aggregating the regional latents of different timesteps

- ❑ Pipelining: Batchify latents of foreground & background prompts in different timesteps into a single batch to maximize throughput.
- ❑ Cache text embeddings to minimize redundancy.
- ❑ Near real-time generation for interactive content creation.

EXPERIMENTS

Main Results

Method	Sampler	FID ↓	IS ↑	$CLIP_I$ ↑	$CLIP_T$ ↑	Time (s) ↓
SD1.5 (512 × 512)						
MultiDiffusion (Ref.)	DDIM [47]	70.93	16.24	24.09	27.55	14.1
MultiDiffusion (MD)	LCM [31]	270.55	2.653	22.53	19.63	1.7
SemanticDraw (Ours)	LCM [31]	93.93	14.12	24.14	24.00	1.3

Method	Sampler	FID ↓	IS ↑	$CLIP_I$ ↑	$CLIP_T$ ↑	Time (s) ↓
SDXL (1024 × 1024)						
MultiDiffusion (Ref.)	DDIM [47]	73.77	16.31	24.16	28.11	50.6
MultiDiffusion (MD)	EulerDiscrete [18]	572.95	1.328	21.02	17.36	4.3
SemanticDraw (Ours)	EulerDiscrete [18]	84.27	15.04	24.19	24.22	3.6

Method	Sampler	FID ↓	IS ↑	$CLIP_I$ ↑	$CLIP_T$ ↑	Time (s) ↓
SD1.5 (512 × 512)						
MultiDiffusion (Ref.)	DDIM [47]	70.93	16.24	24.09	27.55	14.1
MultiDiffusion (MD)	Hyper-SD [40]	168.34	10.12	20.08	15.90	1.7
SemanticDraw (Ours)	Hyper-SD [40]	98.60	14.90	24.48	23.31	1.3

Method	Sampler	FID ↓	IS ↑	$CLIP_I$ ↑	$CLIP_T$ ↑	Time (s) ↓
SD3 (1024 × 1024)						
MultiDiffusion (Ref.)	FlowMatch [9]	166.42	8.517	20.66	16.39	46.3
MultiDiffusion (MD)	FlashFlowMatch [7]	209.36	5.347	19.83	14.48	4.0
SemanticDraw (Ours)	FlashFlowMatch [7]	79.2	17.41	23.59	27.83	3.2

- ❑ Ours enjoy speedup of the acceleration module while maintaining quality.
- ❑ Our method is general; it can be applied to any large-scale diffusion pipeline.

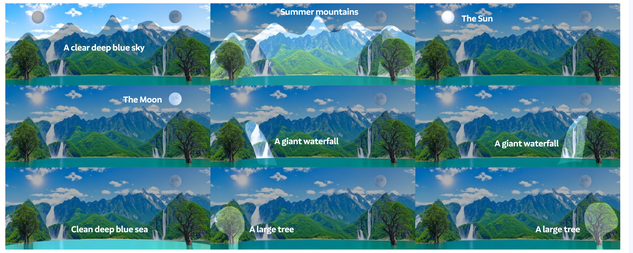
Ablation Study

Method	FID ↓	$CLIP_I$ ↑	$CLIP_T$ ↑
No stabilization	270.55	22.53	19.63
+ Latent pre-averaging	80.64	22.80	26.95
+ Mask-centering bootstrapping	79.54	23.06	**26.72**
+ Quantized masks ($\sigma = 4$)	**78.21**	**23.08**	**26.72**

Method	Throughput (FPS)	Relative Speedup
Baseline [5]	0.0189	×1.0
+ Stable Acceleration	0.183	×9.7
+ Multi-Prompt Stream Batch	1.38	×73.0
+ Tiny AutoEncoder [6]	**1.57**	×**83.1**

- ❑ Multi-prompt stream batch architecture allows us sub-second generation speed; ready to be used in real-time applications.

Mask Consistency Breakdown



- ❑ We gain enhanced mask consistency & harmonization in addition to speedup.

Take Home Message

New wine in new wineskin; New powerful tool (diffusion models) deserves new applications.

Thank you for coming by! Please also have a look at our project page! →

Stars 562