GeoDreamer: Geometry-Guided Diffusion via Implicit Spatial Learning

GeoDreamer: Geometry-Guided Diffusion via Implicit Spatial Learning

Objective

To develop a geometry-aware text-to-image diffusion model that learns spatial structure, semantic coherence, and view-consistent features using DINO-based geometric priors and CLIP-driven textual conditioning   only during training.

At inference time, the model generates high-quality images from text alone, having implicitly learned spatial priors from prior geometry supervision.

Motivation

Diffusion models like Stable Diffusion produce visually stunning results, but often hallucinate geometry or exhibit inconsistent object structures   particularly in few-shot or fine-grained domains.

Your prior NeRF+DINO experiments showed that limited views hinder 3D reconstruction. In this project, we invert the paradigm: instead of learning geometry from views, we use geometry as a teacher to improve 2D generation.

Methodology

Dataset:

- Primary: Oxford Flowers 102 (8,000+ images, 102 fine-grained flower categories)
- Optional: Stanford Cars / CUB-200 for evaluating view and shape consistency

Semantic Conditioning:

- Use CLIPTextModel (ViT-B/16)
- Extract [CLS] token embedding for each class (e.g., "daisy")

Geometry Conditioning (Train-Time Only):

- Use frozen DINOv2-ViT-Small

- Extract patch tokens (excluding CLS)
- Denoted as geometry_tokens (shape: B T D)

Diffusion Pipeline:

- U-Net-based DDPM (UNet2DConditionModel)
- Inputs: x_t, t, encoder_hidden_states = concat(clip_proj, dino_proj)
- Projections align dimensions: clip_proj (512 attn_dim), dino_proj (384 attn_dim)

Training Design:

- Modality Dropout: randomly drop CLIP or DINO embeddings during training
- Classifier-Free Guidance (CFG): train with both condition and null inputs

Evaluation Metrics

- FID: Realism vs. dataset distribution
- CLIPSim: Prompt-image semantic alignment
- LPIPS / SSIM: Structural and view consistency
- Human ranking: Optional perceptual realism rating

Expected Outcomes

- Geometry-aware images from text-only prompts
- Structural consistency and spatial quality improvements vs. baseline DDPM
- Ablation studies comparing DINO / CLIP / both
- Optional: NeRF-based supervision extension

Stretch Goal: NeRF-Guided Extension

Experiment: NeRF as a Geometry Teacher

- Train a small NeRF on 10 20 views per class
- Extract scene latent or volume features

- Pass to U-Net as nerf_proj tokens
- Train U-Net with CLIP + DINO + NeRF conditioning

Experimental Phases

----

1. Baseline pipeline: clip_encoder.py, dino_encoder.py, unet_with_proj.py, train_baseline.py
2. Conditioning injection: geometry_aware_unet.py
3. Modality dropout & CFG: train_dropout_cfg.py
4. Sampling & visualization: sample_images.py
5. Metric evaluation: eval_metrics.py
6. Ablation study: encoder_hidden_states manipulation
7. NeRF supervision: nerf_teacher.py, nerf_features.py, LoRA adapter
8. Distillation (optional): distill_student.py

Directory Structure (Suggested)

```
geodreamer/
  data/
      flowers/
  models/
      clip_encoder.py
      dino_encoder.py
      geometry_aware_unet.py
      nerf_teacher.py
  training/
      train_baseline.py
      train_dropout_cfg.py
      distill_student.py
  inference/
      sample_images.py
      generate_comparisons.py
  evaluation/
      eval_metrics.py
      ablation_plots.ipynb
  checkpoints/
```

Tools & Resources

- Training: Kaggle Notebooks (A100/T4 GPU)
- Inference & Visuals: MacBook (M4 Pro)
- Libraries: PyTorch, transformers, diffusers, DINOv2, CLIP, nerfstudio, taming-transformers, clean-fid, lpips

Final Notes

This proposal includes a practical, testable architecture with text-only generation at inference, modular design, research-oriented extensions, and a structured, phase-driven execution plan for rapid iteration and publication readiness.