Experiment Log

CogVideoX-2b: Text-to-Video Generation & Hyperparameter Ablation Studies

Claudio Dragotta

1. Objective & Setup

This project explores text-to-video generation using CogVideoX-2b, a 2B parameter diffusion model by THUDM (Tsinghua University). We demonstrate inference across 9 diverse prompts and run controlled ablation studies on 4 hyperparameters, varying one at a time on a fixed prompt (panda_guitar) to isolate each parameter's effect.

Environment: WSL2 Ubuntu, NVIDIA GPU with CUDA 12.x, Python 3.12, PyTorch, Hugging Face Diffusers >= 0.30.1. Model loaded in float16 with sequential CPU offload + VAE tiling/slicing for ~15 GB VRAM. A Colab notebook (T4/A100) is provided.

2. Pipeline Design

The script (run_cogvideox.py) uses a file-based prompt system: each prompt is a .txt file in prompts/. Adding a new scenario requires zero code changes. All hyperparameters are CLI arguments. A batch script (run_experiments.sh) generates all 20 experiment videos organized in subfolders (baseline/, exp1_guidance_scale/, exp2_steps/, exp3_seed/, exp4_frames/) with skip logic for safe resumption.

Parameter	Default	Role
--steps	50	Number of diffusion inference steps
--guidance	6.0	Classifier-free guidance scale
--frames	49	Number of video frames to generate
--seed	42	Random seed for reproducibility

3. Experiments & Results

Baseline: 9 videos generated with default parameters (steps=50, GS=6, seed=42, frames=49). All produced successfully across realistic, fantastical, and cinematic scenarios. Each video takes ~10-15 minutes to generate.

Exp 1: guidance_scale (1 vs 6 vs 12)

- **GS=1:** Panda shape present but severely distorted, no guitar visible. Background noisy.
- **GS=6 (default):** Excellent quality. Panda with red jacket, guitar, bamboo forest all clearly rendered. Smooth motion, natural colors. Best overall result.
- **GS=12:** Sharper and more vivid than GS=6, but colors oversaturated and appearance more '3D rendered' / artificial. Still usable but less natural.

Exp 2: num_inference_steps (10 vs 25 vs 50)

- **10 steps:** COMPLETE FAILURE -- dark black blob on green background with scan-line artifacts. No recognizable content. File is ironically the largest (1.1 MB: noise compresses poorly).
- **25 steps:** Surprisingly near-perfect quality, very close to 50 steps at roughly half the time.
- **50 steps:** Full quality baseline. Maximum detail and smooth motion. File size 767 KB.

Exp 3: seed (42 vs 123 vs 999)

Each seed produces a genuinely different composition -- different camera angles, lighting, and background elements -- not just minor noise variations. Quality remains consistent. Zero impact on generation time.

Exp 4: num_frames (25 vs 49)

Per-frame quality is identical; only video duration changes (3s vs 6s at 8 FPS). File size scales linearly (337 KB vs 767 KB). 49 frames allows complete motion cycles; 25 frames is useful for quick previews.

4. Analysis & Conclusions

Rank	Parameter	Affects	Time	Recommendation
1	guidance_scale	Quality & adherence	None	Keep at 6
2	num_inference_steps	Quality & speed	Linear	50 final, 25 preview
3	seed	Composition & style	None	Try 3-5, pick best
4	num_frames	Duration & VRAM	Proportional	49 final, 25 preview

- **Key discovery:** there is a critical threshold at ~20-25 inference steps below which the model completely fails. The jump from 10 to 25 steps is the largest quality leap in all experiments, disproving the assumption that quality degrades linearly.
- **File size as quality proxy:** noisy 10-step video = 1.1 MB (high entropy), clean 50-step = 767 KB. Coherent images compress better than noise.
- **Practical workflow:** (1) write prompt, (2) preview with steps=25, (3) try different seeds if needed, (4) run best seed at steps=50 for final output.

CogVideoX-2b generates coherent videos from complex prompts in ~10-15 min on consumer hardware. The defaults (GS=6, steps=50) are well calibrated. Memory optimizations (CPU offload, VAE tiling/slicing) are essential for ~15 GB VRAM. Future work: CogVideoX-5b, LoRA fine-tuning, quantitative FVD/CLIP evaluation.