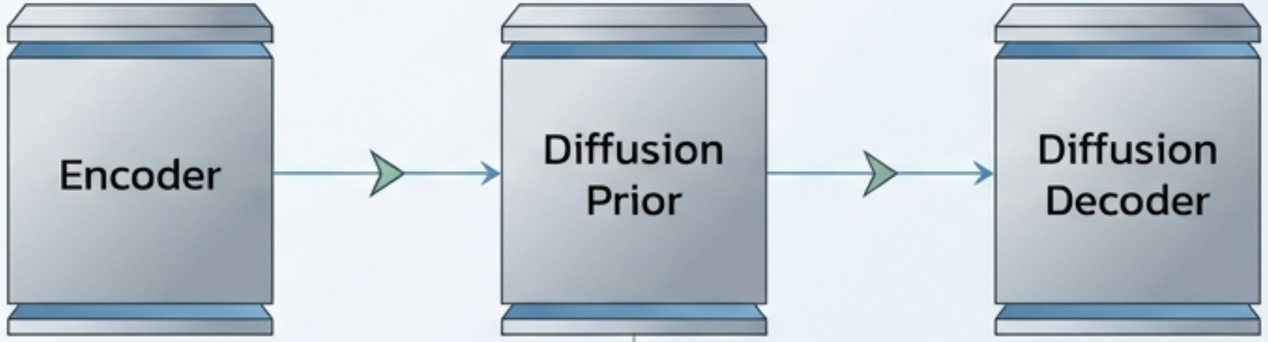
Google DeepMind

Unified Latents (UL): โมเดลพื้นฐานยุคใหม่จาก Google DeepMind

วิธีการเทรน Latents อย่างเป็นระบบ
เพื่อประสิทธิภาพระดับ
State-of-the-Art

Unified Latents (UL) คืออะไร?

UL คือเฟรมเวิร์กใหม่สำหรับการเรียนรู้ Latent Representations โดยใช้ Diffusion Model ทั้งในการทำ Regularization (Prior) และ Decoding



ประสิทธิภาพระดับ State-of-the-Art (SOTA):

FID 1.4

ImageNet-512: แซงหน้าทุกโมเดลพร้อมประหยัด Compute มหาศาล

FVD 1.3

Kinetics-600: สร้างสถิติใหม่ของโลก

- หมดปัญหาการเดาสุ่มในการจูนพารามิเตอร์ ด้วยสมการที่ควบคุมปริมาณข้อมูล (Bitrate) ได้อย่างแม่นยำ

NotebookLM

ปัญหาคลาสสิกของ Latent Diffusion Models (LDMs)

ความขัดแย้งของข้อมูล (The Latent Trade-Off):

Latents ที่เรียนรู้ง่าย (Low Information)

บีบอัดข้อมูลมากเกินไป ทำให้สูญเสียรายละเอียดและโครงสร้างความถี่สูง (Poor PSNR)

Latents ที่มีรายละเอียดสูง (High Information)

เก็บรายละเอียดภาพได้ครบถ้วน แต่ทำให้ Base Model ทำงานหนักและเรียนรู้ได้ยาก

บทสรุป: ยิ่ง Latents เรียนรู้ง่ายเท่าไร คุณภาพในการสร้างภาพ (Generation Quality) ก็จะยิ่งดีขึ้นเท่านั้น หากเรารักษาความหนาแน่นของข้อมูลไว้ได้

NotebookLM

ทำไม Standard VAEs ถึงไม่ใช่คำตอบ?

⚠️ **สูญเสีย High-Frequency**: การใช้ Regularized Autoencoders แบบเดิม มักทำให้ภาพเกิด Artifacts หรือสูญเสียรายละเอียดเล็กๆ น้อยๆ ไปในการสร้างภาพใหม่

ขาด Likelihood-based loss

Decoder ของ VAE ทั่วไปขาดระบบสูญเสีย (Loss) ที่อิงตามความน่าจะเป็น

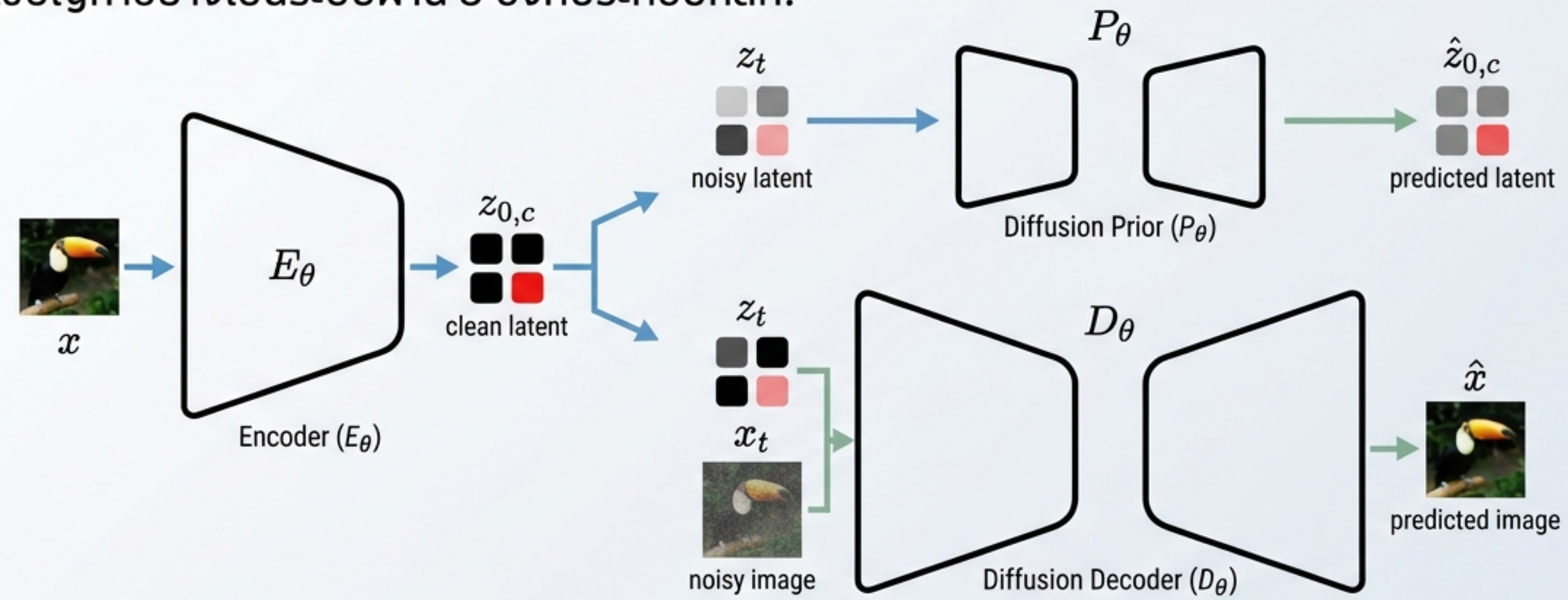
ภาระการปรับจูน (Manual Tuning)

น้ำหนักของ KL Penalty ต้องถูกปรับตั้งค่าด้วยมือ (Manual) ทำให้ประเมินปริมาณข้อมูล (Information Content) ของ Latent ได้ยาก

NotebookLM

สถาปัตยกรรม Unified Latents (UL)

การแก้ไขปัญหาอย่างเป็นระบบผ่าน 3 องค์ประกอบหลัก:



1. Encoder (E_θ)
แปลงภาพให้เป็น Latent โดยกำหนดค่า Gaussian Noise ให้คงที่

2. Diffusion Prior (P_θ)
โมเดลสำหรับกะประมาณและจำกัดขอบเขตของ Information Content โดยอิงกับระดับ Noise ต่ำสุด

3. Diffusion Decoder (D_θ)
โมเดลสำหรับถอดรหัสภาพกลับโดยใช้ Reweighted ELBO loss

NotebookLM

นวัตกรรมที่ 1: Fixed Encoding Noise

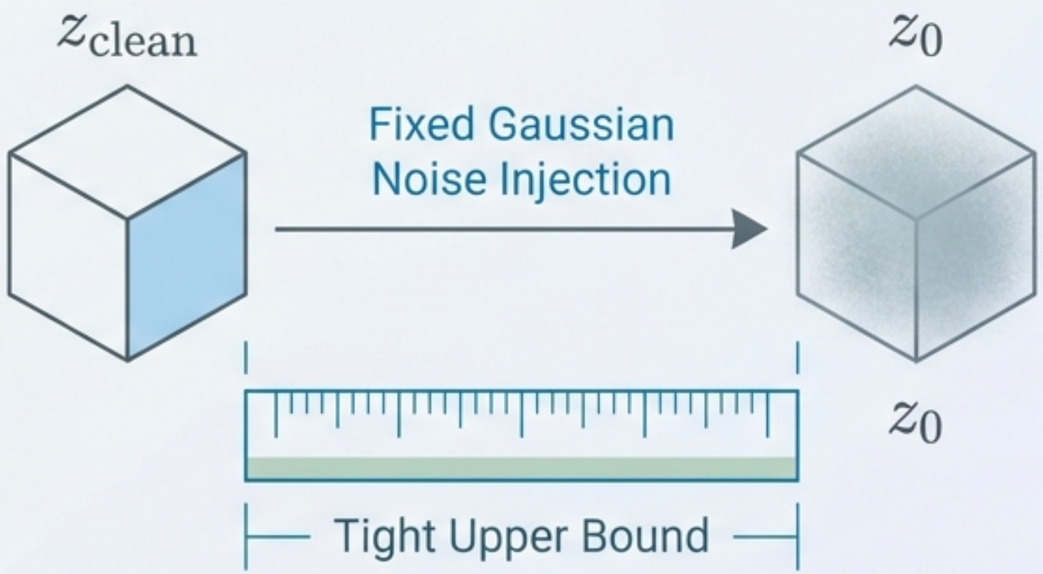
จากเดิม: VAE แบบดั้งเดิมปล่อยให้โมเดล (เช่น Stable Diffusion) เรียนรู้และสุ่มระดับ Noise เอง ซึ่งมักก่อให้เกิดความไม่เสถียร (Instability)

แนวทางของ UL:

– ใช้ Deterministic Latent เรียบๆ แล้วเพิ่ม Noise เข้าไปจนถึงเวลา $t = 0$ ที่กำหนดไว้ (Fixed Amount)

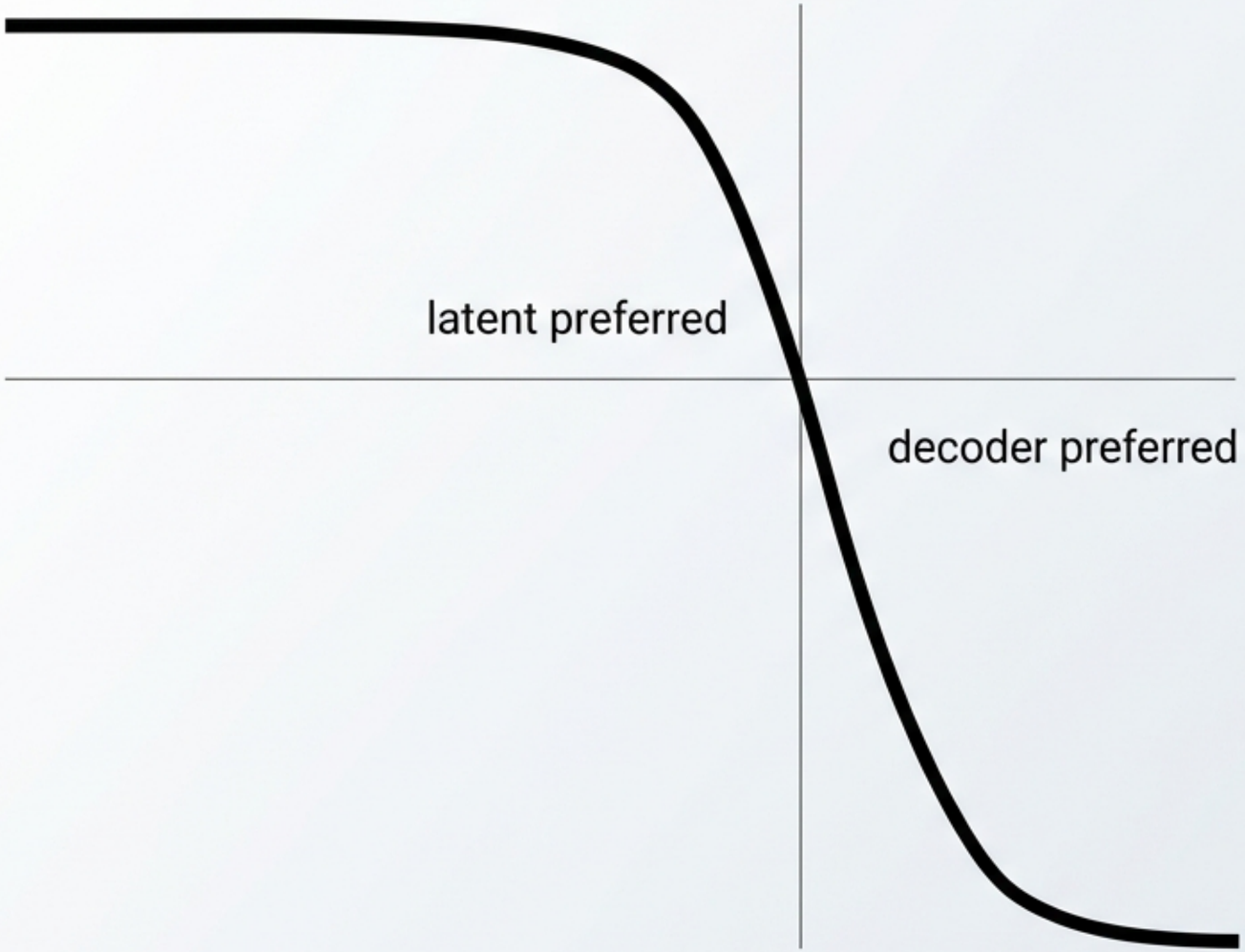
– เชื่อมโยง Noise ของ Encoder เข้ากับระดับความแม่นยำ (Precision) สูงสุดของ Diffusion Prior

ผลลัพธ์: สร้างขอบเขต **(Tight Upper Bound)** ของ **Latent Bitrate** ที่คำนวณและตีความได้อย่างชัดเจน



NotebookLM

นวัตกรรมที่ 2: The Diffusion Decoder



Decoder ทำหน้าที่ทำงานในระดับ Pixel Space โดย Condition กับทั้งข้อมูลที่มี Noise (x_t) และ Latent (z_0)

Reweighted ELBO Loss:
– ใช้วิธีลดน้ำหนัก (Discounting) ความสำคัญในช่วงระดับ Noise ที่ต่ำ

เคล็ดลับ:
การทำเช่นนี้เป็นการ “บังคับ” ให้ Decoder ต้องรับผิดชอบการสร้างรายละเอียดความถี่สูง (High-frequency features) เสมอ เพราะมีต้นทุนของข้อมูล (Cost per bit) ที่ถูกกว่า

NotebookLM

The Tuning Knob: ควบคุมสมดุลด้วย Loss Factor (c_{lf})

แม้จะมีการให้น้ำหนักที่เท่ากัน
แต่ปัญหา "Posterior Collapse"
(โมเดลเพิกเฉยต่อ Latent)
ก็ยังเกิดขึ้นได้หาก Decoder
ทรงพลังเกินไป

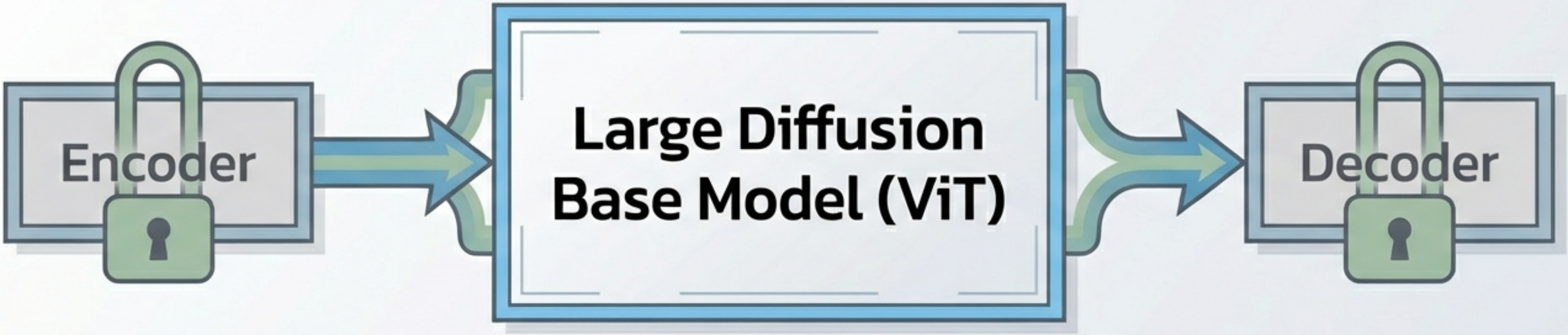
การแก้ปัญหาด้วย c_{lf}:

- UL เพิ่มน้ำหนักให้กับ Decoder Loss อย่างชัดเจน (มักตั้งค่า c_{lf} ระหว่าง 1.3 ถึง 1.7)
- เป็นพารามิเตอร์ที่เรียบง่าย แต่มีประสิทธิภาพในการควบคุมปริมาณข้อมูล (Information) ใน Latent
- Latent ที่มีข้อมูลมาก = ภาพสมจริง (Reconstruction) มากขึ้น แต่ Base Model ต้องรับภาระมากขึ้น

NotebookLM

ขั้นตอนที่ 2: การเทรน The Base Model

เมื่อได้คุณสมบัติ Latent ที่สมบูรณ์จาก Stage 1 แล้ว กระบวนการจะเข้าสู่การสเกล



Freeze the Edges: ล็อก (Freeze) ค่าน้ำหนักของ Encoder และ Decoder ไว้

Train the Core: เทรน Diffusion Prior Model (Base Model) ขนาดใหญ่ด้วย Latents ที่ได้ สามารถขยายขนาดโมเดลและ Batch Size ได้มหาศาลอย่างมีประสิทธิภาพ

NotebookLM

บทพิสูจน์ระดับ SOTA บน ImageNet-512

Training cost (x-axis) vs. FID (y-axis)

UL (ours) – ต่ำสุดและดีที่สุด

ประสิทธิภาพที่เหนือกว่า:

- UL ทำลายสถิติเดิม แซงหน้า DiT, EDM2 และ Stable Diffusion (SD) Baselines

ความคุ้มค่าระดับสุดยอด:

- ทำ FID ได้ต่ำถึง 1.4
- ใช้จำนวน FLOPs ในการเทรนน้อยกว่าโมเดลที่พึ่งพา SD Latents อย่างมีนัยสำคัญ
- ประหยัดพลังงานประมวลผล (Compute) แต่ได้ผลลัพธ์การสร้างภาพที่ดีที่สุด

NotebookLM

การขยายสเกลสู่ Text-to-Image

เมื่อนำไปเทรนบนชุดข้อมูล Text-to-Image ขนาดใหญ่ UL แสดงให้เห็นถึงคุณภาพที่ก้าวกระโดด

เหนือกว่า Pixel Diffusion & SD: ให้คุณภาพการสร้างภาพภาพรวม (gFID) ที่ดีเยี่ยม

ความสอดคล้องกับข้อความ (CLIP Alignment) เหนือกว่าอย่างชัดเจน แม้ Base Model จะมีขนาดเล็ก

NotebookLM

ความทนทานต่อรูปแบบ Latent (Robustness)

ปัญหาดั้งเดิมของ LDM คือการตัดสินใจเลือกจำนวน Channel และการทำ Downsampling

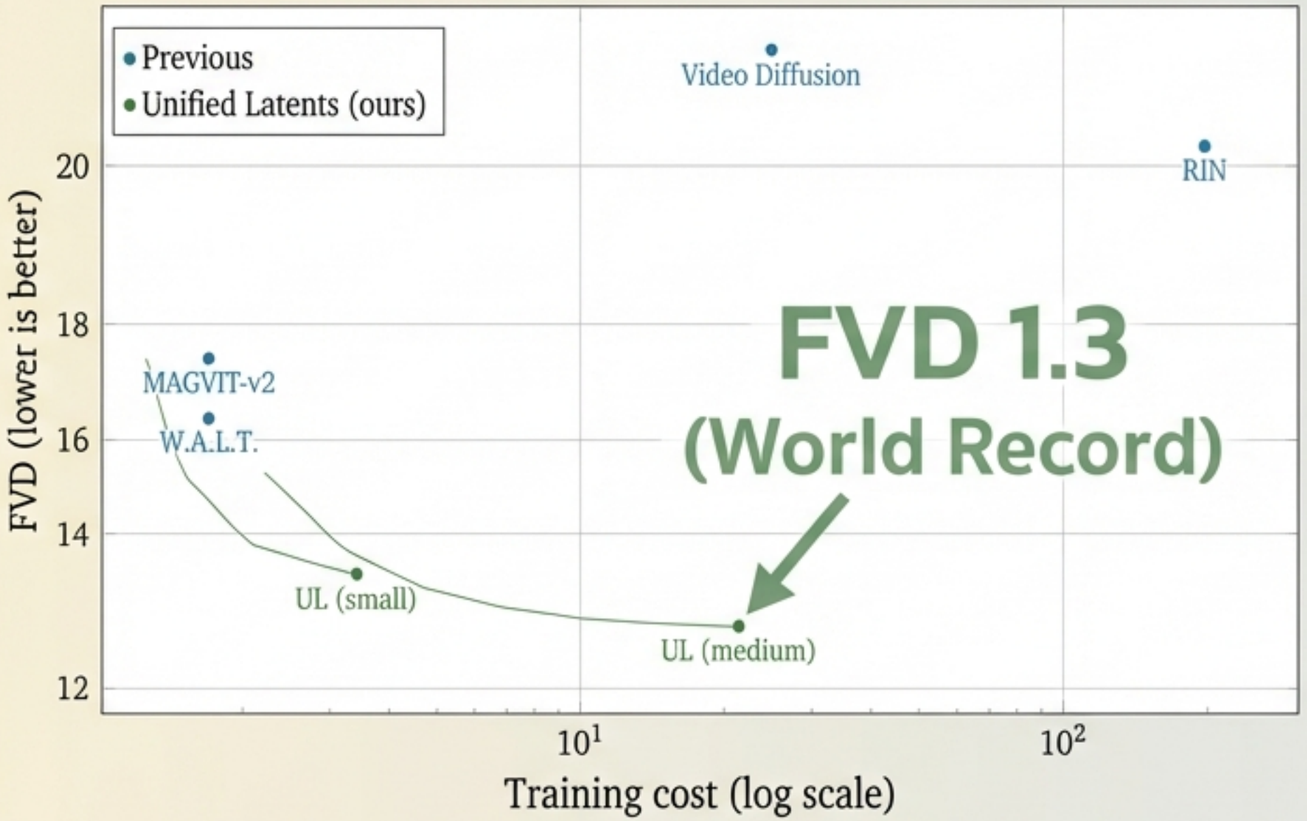
Latent Shape	rFID@50K	gFID@50K
64x64x32	0.40	2.12
32x32x32	0.41	1.63
16x16x32	1.41	1.74

ข้อมูลพิสูจน์แล้วว่า UL แทบไม่สะทกสะท้าน:

- Latent Channels: คงประสิทธิภาพ rFID และ gFID ได้อย่างนิ่งสนิท ไม่ว่าจะใช้ 16, 32 หรือ 64 Channels
- Spatial Downsampling: รองรับการบีบอัดตั้งแต่ 8x ถึง 32x โดยคุณภาพแทบไม่ตกหล่น
- ขจัดความยุ่งยากในการทดลองสถาปัตยกรรม (Architectural Sweeps) สำหรับทีมวิจัย

NotebookLM

สถิติโลกใหม่ด้าน Video Generation (Kinetics-600)

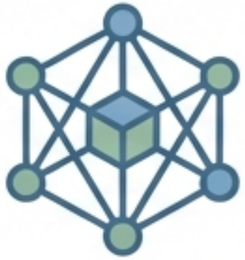
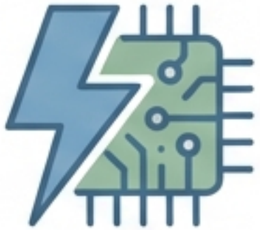
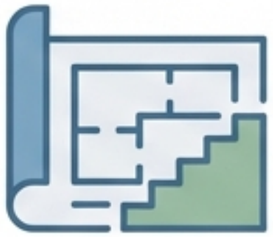


- UL ไม่ได้จำกัดอยู่แค่ภาพนิ่ง แต่เป็นรากฐานที่สมบูรณ์แบบสำหรับวิดีโอ

- The World Record:
  - โมเดล UL ขนาดกลาง (Medium) บรรลุ FVD ระดับ State-of-the-Art ที่ 1.3
  - โมเดล UL ขนาดเล็ก (Small) ยังยังสามารถทำ FVD ได้ที่ 1.7
  - เอาชนะคู่แข่งระดับท็อป (MAGVIT-v2, W.A.L.T., Video Diffusion) โดยใช้ Compute เพียงเศษเสี้ยว

NotebookLM

สรุปประเด็นสำคัญและทิศทางในอนาคต

		
ทลายข้อจำกัด UL แกะรหัส Trade-off ระหว่าง ‘การบีบอัดข้อมูล’ และ ‘การสร้างภาพ’ ได้อย่างสมบูรณ์ ด้วยการผสมผสาน Diffusion Prior และ Decoder	**ประสิทธิภาพสูงสุด** ให้ SOTA Metrics อย่างชัดเจน (FID 1.4 / FVD 1.3) คุ้มค่าในทุก FLOPs ที่ใช้เทรน	**มาตรฐานใหม่ของวงการ** UL มอบ Hyper-parameters ที่ควบคุมได้อย่างโปร่งใส กลายเป็นพิมพ์เขียว (Blueprint) ใหม่ที่ทรงพลังสำหรับการสร้าง Foundation Models ขนาดใหญ่ในอนาคต

NotebookLM

ศักยภาพเหนือขีดจำกัดของ Unified Latents

* ภาพสร้างจากการป้อนข้อความ (Zero-shot Text-to-Image) โดยโมเดล UL พื้นฐาน ไม่มีการปรับแต่งเพิ่มเติม

NotebookLM