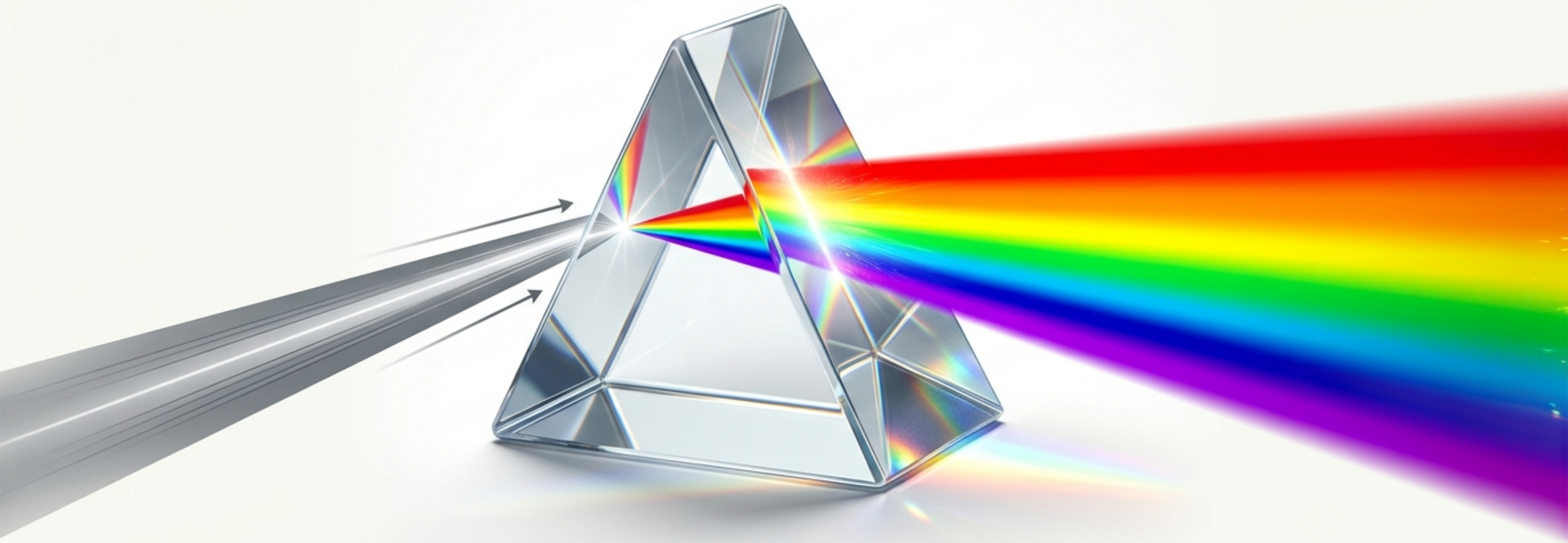
สมมติฐานปริซึม: กำหนดกระบวนทัศน์ใหม่ของ Generative AI

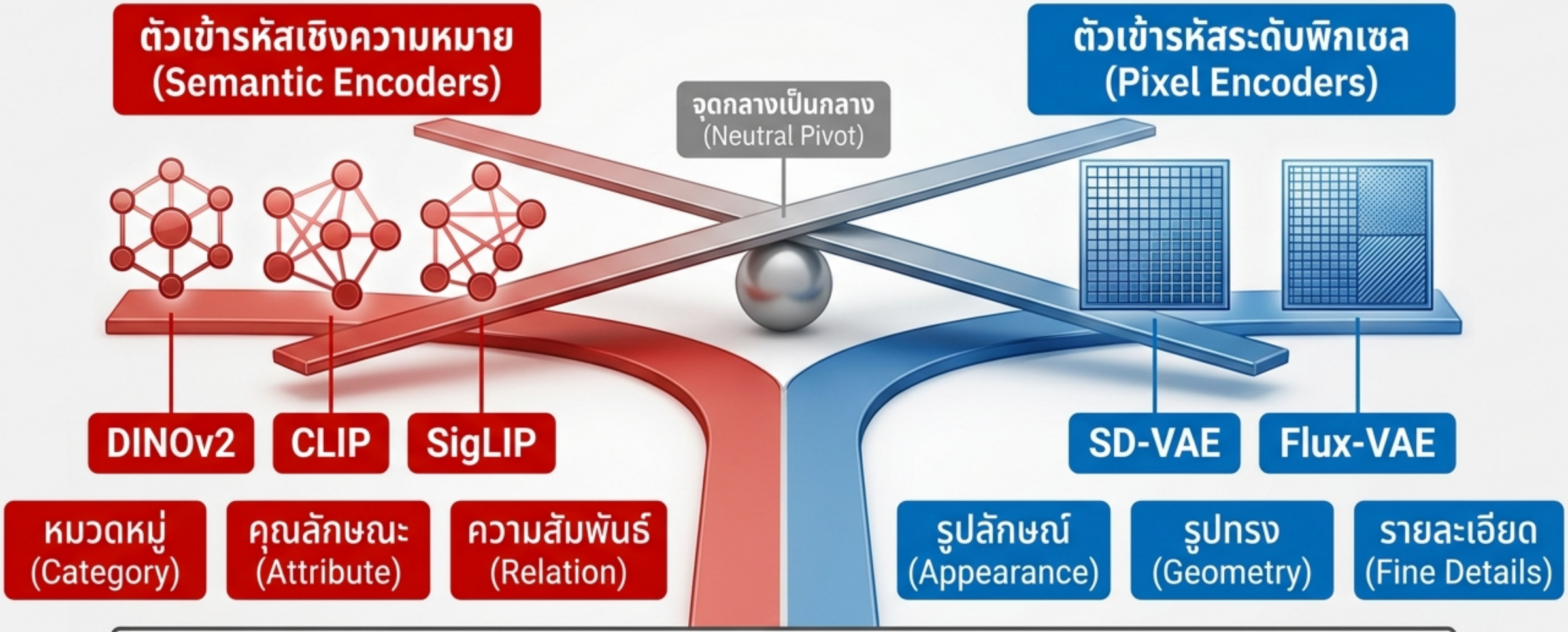
ผสานความเข้าใจบริบท (Semantics) และความคมชัด (Pixels) ไว้ในพื้นที่แฝงเดียวด้วย Unified Autoencoding (UAE)



Weichen Fan, Haiwen Diao, Quan Wang, Dahua Lin, Ziwei Liu

NotebookLM

ความขัดแย้งของตัวแบบรากฐาน (Foundation Models)



ปัจจุบัน AI ต้องเลือกระหว่าง ‘ความเข้าใจความหมาย’ หรือ ‘ความสมจริงของภาพ’
การพยายามใช้ทั้งสองระบบพร้อมกันทำให้เกิดความขัดแย้งของข้อมูลและใช้ทรัพยากรมหาศาล

NotebookLM

กระบวนทัศน์ของระบบการนำเสนอข้อมูล (Representation Paradigm)

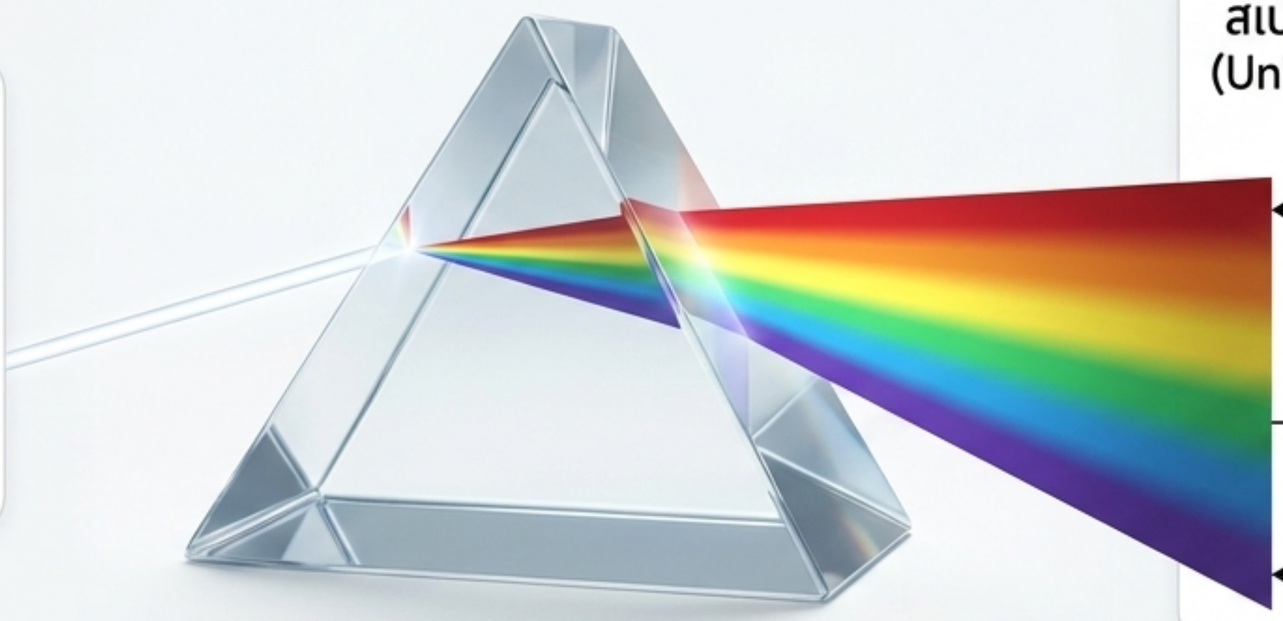
	Semantic Encoders (เช่น DINOv2)	Pixel Encoders (เช่น SD-VAE)	**Unified Autoencoding (UAE)**
โฟกัสหลัก (Primary Focus)	นามธรรมและโครงสร้าง (Abstraction)	รายละเอียดและพื้นผิว (Texture)	ความเข้าใจผสานความคมชัด (Unified)
ย่านความถี่หลัก (Dominant Frequency)	ความถี่ต่ำ (Low-Frequency)	ความถี่สูง (High-Frequency)	เต็มสเปกตรัม (Full Spectrum)
ความสมจริงของภาพ (Visual Fidelity)	ต่ำ (เบลอ/ผิดเพี้ยน)	สูงมาก (คมชัด)	สูงเทียบเท่า Pixel Encoders
ความเข้าใจบริบท (Semantic Grasp)	ลึกซึ้ง	ผิวเผิน	ลึกซึ้งเทียบเท่า Semantic Encoders

NotebookLM

สมมติฐานปริซึม
(The Prism Hypothesis)

ข้อมูลในโลกความเป็นจริงเป็นเสมือนแสงที่
สามารถแยกส่วนประกอบตามความถี่ได้

มุมมองตามธรรมชาติ
(Natural Inputs)



สเปกตรัมแฝงแบบรวมศูนย์
(Unified Latent Spectrum)

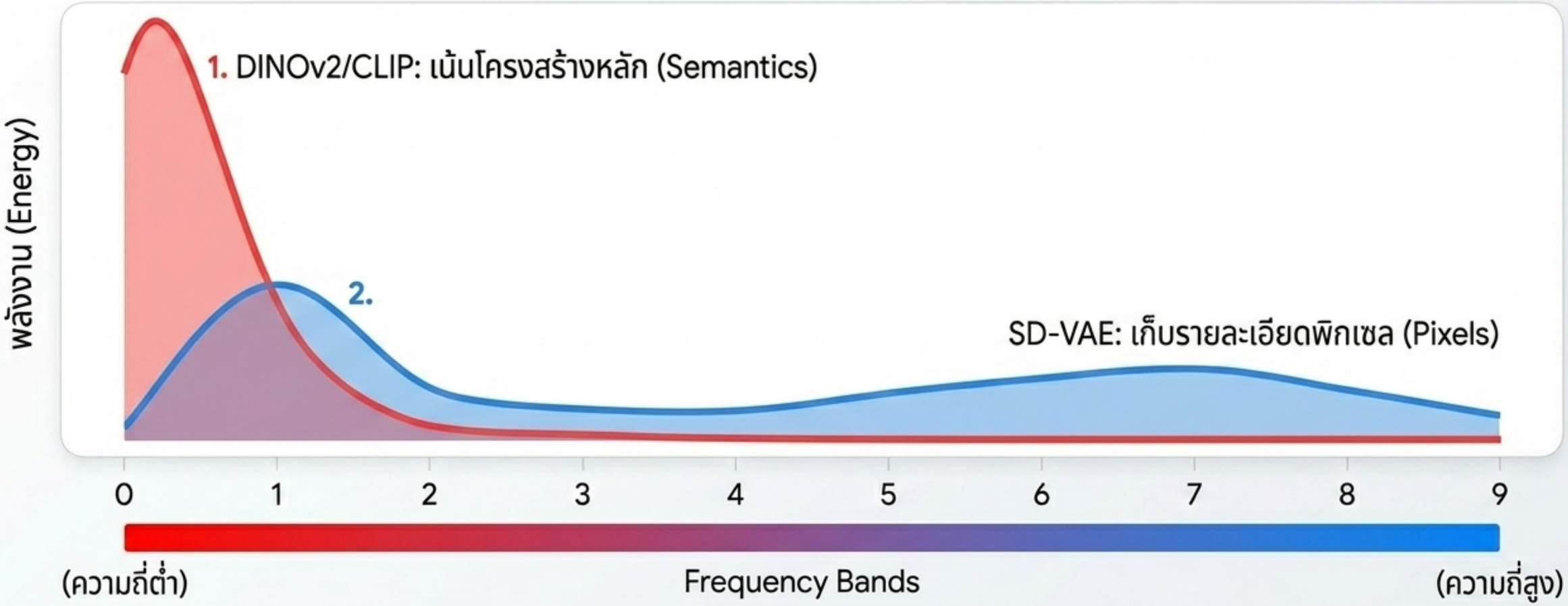
← **ความถี่ต่ำ:** โครงสร้างหลัก, ความหมาย, หมวดหมู่

ความถี่สูง: รายละเอียด, พื้นผิว, ความคมชัด

โมดอลิตี้ (Modality) ที่ต่างกัน
แท้จริงแล้วคือมุมมองที่แตกต่างกันบนสเปกตรัมเดียวกัน

NotebookLM

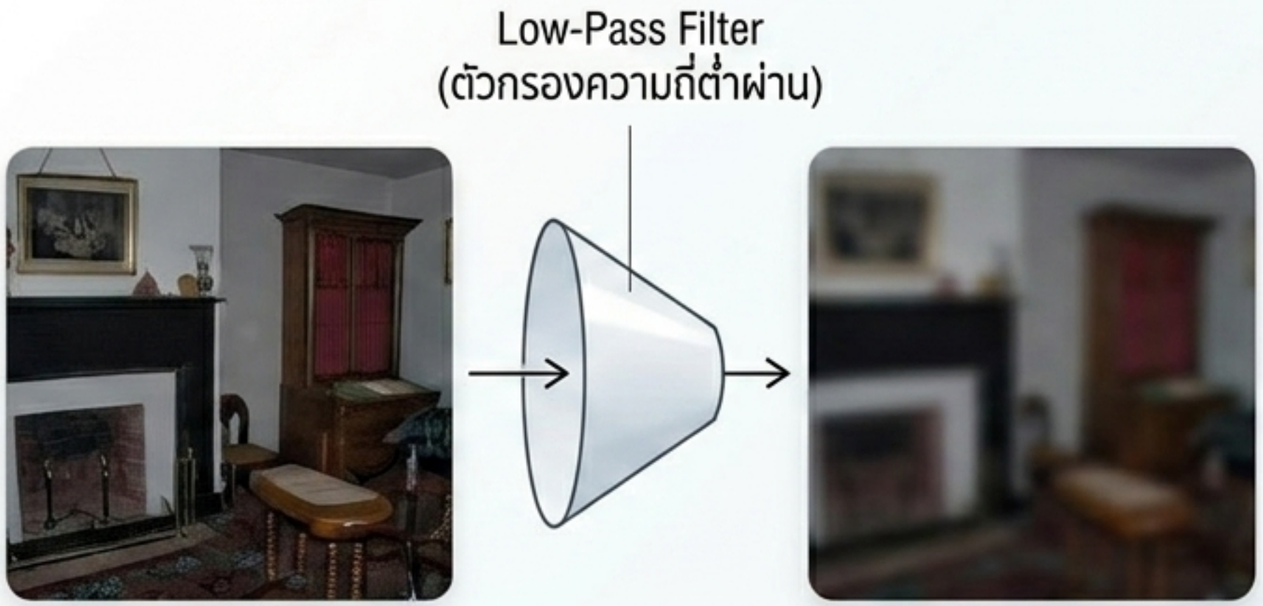
บทพิสูจน์ที่ 1: ความเอนเอียงทางสเปกตรัม (Spectral Bias)



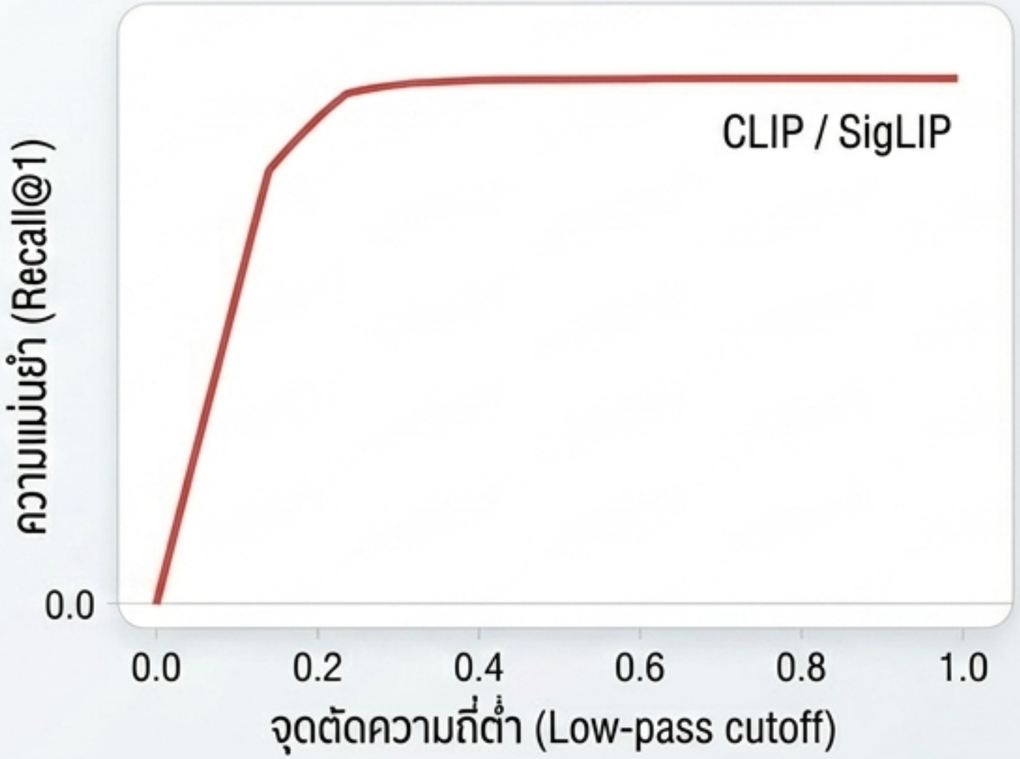
ตัวเข้ารหัสเชิงความหมายกระจุกตัวอยู่ที่ความถี่ต่ำอย่างชัดเจน
ในขณะที่ตัวเข้ารหัสพิกเซลยังคงรักษาพลังงานในย่านความถี่สูงไว้

NotebookLM

บทพิสูจน์ที่ 2: ความหมายแฝงอยู่ในคลื่นความถี่ต่ำ



ภาพที่ถูกตัดความถี่สูงออก
(Low-Pass Filtered)



แม้จะตัดความถี่สูง (รายละเอียดภาพ) ทิ้งไปจนเกือบหมด
โมเดลก็ยังจับคู่ภาพกับข้อความได้อย่างแม่นยำ

โครงสร้างความหมายของภาพ ไม่ต้องการรายละเอียดระดับพิกเซล

NotebookLM

ทางออก: Unified Autoencoding (UAE)

=

+

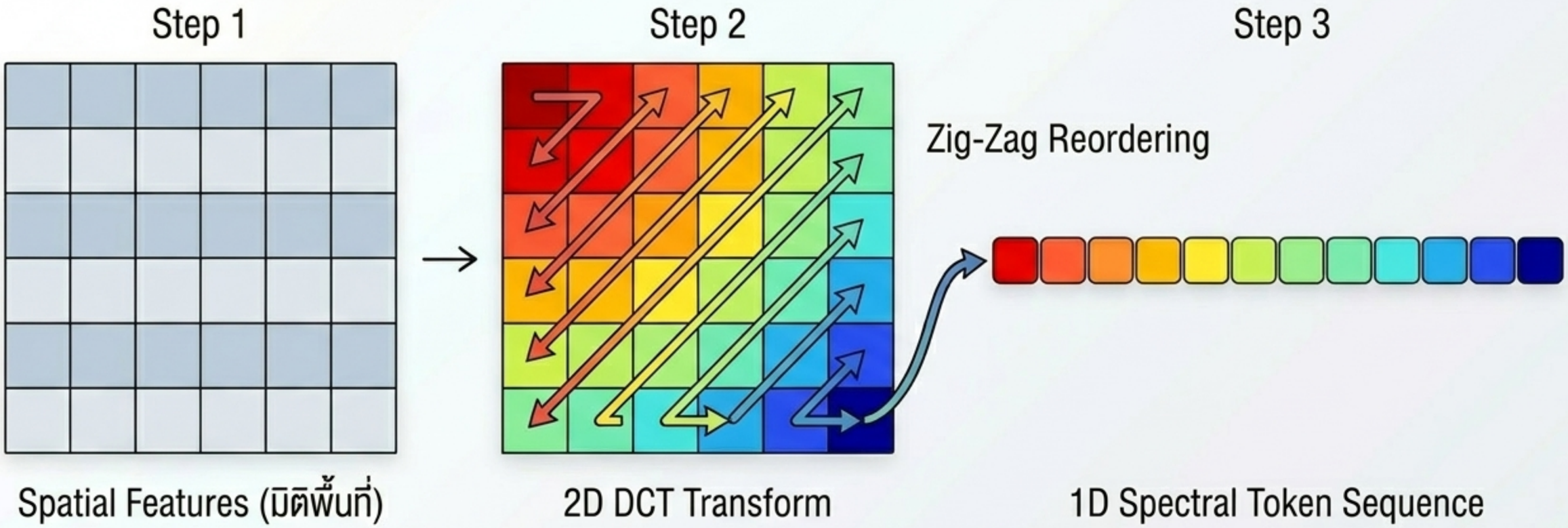
(ภาพป้อนเข้า) = (ฐานความหมายที่ใช้ร่วมกัน) + (รายละเอียดพิกเซลส่วนเพิ่ม)

Step 1: ดึงคุณลักษณะสำคัญด้วยตัวเข้ารหัสเดิม (DINOv2) → Step 2: แยกสเปกตรัมความถี่ออกจากกัน → Step 3: บังคับให้ความถี่ต่ำจอิงตามความหมาย และให้ความถี่สูงจัดการพื้นผิวภาพ

รวมจุดเด่นของทั้งสองโลกไว้ใน Latent Space เดียว

NotebookLM

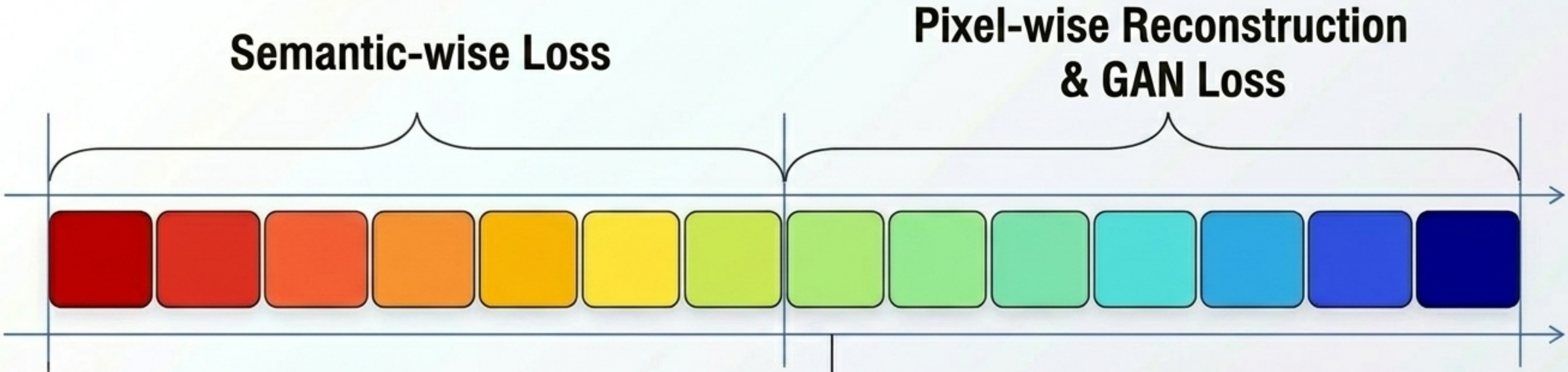
เบื้องหลังการทำงาน: แปลงมิติภาพสู่คลื่นความถี่



- 2D DCT (Discrete Cosine Transform) ทำหน้าที่แยกย่านความถี่ของภาพอย่างเป็นระบบ
- Zig-Zag Reordering นำข้อมูลมาจัดเรียงเป็นโทเค็น 1 มิติ จากความถี่ต่ำสุดไปสู่สูงสุดอย่างเป็นลำดับ

NotebookLM

เป้าหมายคู่ขนาน (Dual Objective Training)



บังคับให้โครงสร้างหลักตรงกับความหมาย
ของ Teacher Model (DINOv2)

ปล่อยอิสระให้โมเดลเรียนรู้การสร้าง
รายละเอียดพิกเซลที่คมชัดและสมจริง

การกำหนดขอบเขตความถี่ที่ชัดเจน ทำให้โมเดลไม่เกิดการขัดแย้ง
กันเองระหว่างการเรียนรู้ ‘ความเข้าใจ’ และ ‘ความคมชัด’

NotebookLM

กุญแจสำคัญ: Frequency Masked Prediction

1. ข้อมูลเต็มสเปกตรัม (Full Spectrum) ⟶

2. ซ่อนความถี่สูสูง (Masking) ⟶

3. คาดเดาและสร้างใหม่ (Prediction)

กลไกการทำงาน:

- ระหว่างการฝึกสอน จะทำการซ่อนข้อมูลความถี่สูงบางส่วนไว้
- บังคับให้ Decoder ต้อง ‘คาดเดา’ รายละเอียดพื้นผิวจากโครงสร้างความหมายหลักที่เหลืออยู่
- ผลลัพธ์: โมเดลเข้าใจความเชื่อมโยงระหว่างนามธรรมและภาพจริงได้อย่างลึกซึ้ง

NotebookLM

ความยืดหยุ่นในการสร้างสรรค์ (Generative Modeling)

UAE Tokens

Spatial-wise Modeling (สร้างภาพจากมิติพื้นที่)

Inverse DCT → 2D Spatial Feature → Diffusion Transformer

วิธีมาตรฐาน เข้ากันได้กับสถาปัตยกรรมเดิมได้ทันที

Frequency-wise Modeling (สร้างภาพตามความถี่)

Diffusion Transformer (Progressive) → Inverse DCT

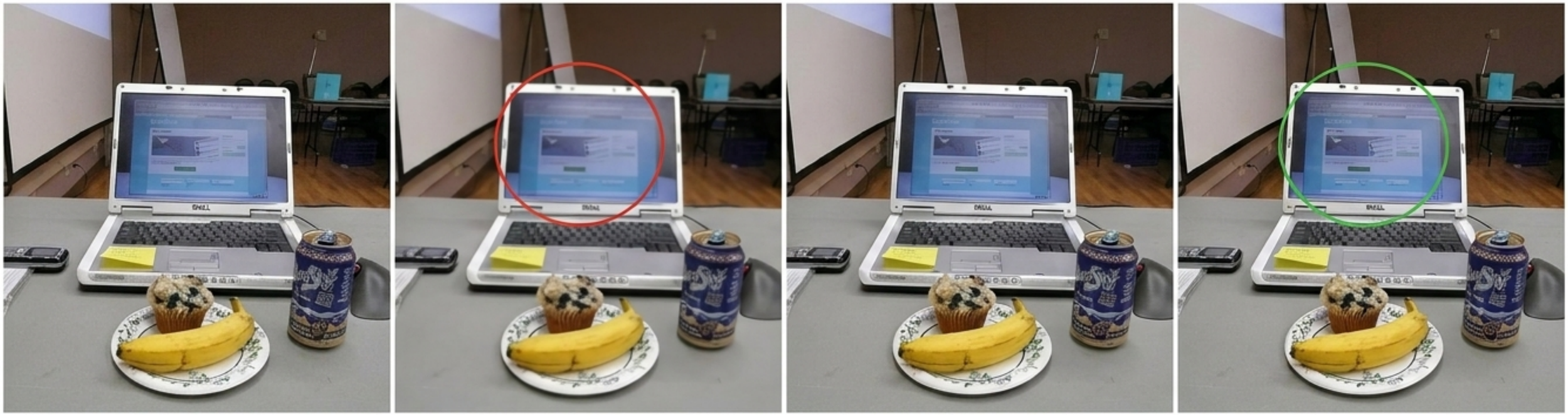
สร้างจากโครงร่างนามธรรม ไปสู่รายละเอียดพิกเซล (Coarse-to-fine)

วิธีใหม่ ประสิทธิภาพสูง ช่วยลดระยะเวลาและทรัพยากรการเทรน

UAE รองรับการทำงาน**ทั้งสองรูปแบบ**
มอบอิสระในการออกแบบสถาปัตยกรรม Generative AI ในอนาคต

NotebookLM

ผลลัพธ์: การสร้างภาพใหม่ที่สมบูรณ์แบบ



ภาพต้นฉบับ (Original)

RAE (DINOv2)
สูญเสียความคมชัด

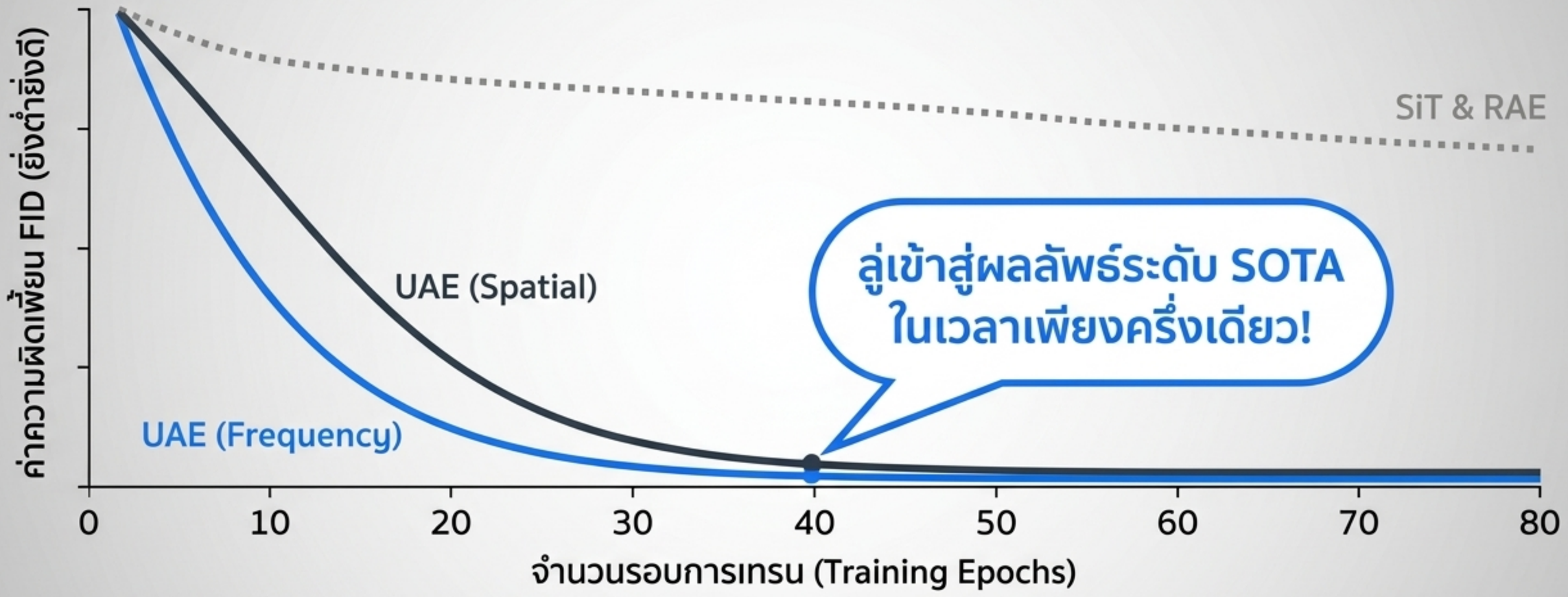
FLUX-VAE
ชัดเจน แต่ใช้ทรัพยากรประมวลผลสูงมาก

UAE (DINOv2) - **ของเรา**
ชัดเจนสมบูรณ์แบบ
ควบคู่กับความเข้าใจบริบท

UAE รักษาขอบที่คมชัด รายละเอียดของพื้นผิว และตัวอักษรขนาดเล็กได้อย่างไร้ที่ติ แก้ปัญหาสีเพี้ยนและภาพเบลอใน RAE

NotebookLM

ความเร็วในการเทรนที่ก้าวกระโดด (Generative Efficiency)



UAE (โดยเฉพาะแบบ Frequency-wise) ลู่เข้าสู่ผลลัพธ์ที่ดีที่สุดเร็วกว่าโมเดลคู่แข่งอย่างมีนัยสำคัญ ร่นเวลาการเทรนลงกว่าครึ่ง โดยให้คุณภาพ FID ที่ดีกว่าที่ระดับ 1.52

NotebookLM

การลดการใช้ทรัพยากรระดับปฏิวัติวงการ

4x

GFLOPs Reduction

ลดปริมาณการประมวลผลต่อภาพลงถึง 4 เท่า
(จาก 313.84 เหลือเพียง 78.80)

2x

Latency Drop

ความเร็วในการทำงาน (Inference)
เพิ่มขึ้นกว่าสองเท่า (8.82ms ลดเหลือ 4.05ms)

0%

Quality Loss

รักษาความแม่นยำด้านโครงสร้าง (PSNR/SSIM)
และคุณภาพการจัดหมวดหมู่เทียบเท่าต้นฉบับ

ด้วยการทำนายผลเฉพาะโทเค็นความถี่ต่ำ (64 Tokens) ร่วมกับ
Frequency-wise modeling ช่วยลดภาระการคำนวณของโมเดล Diffusion ได้มหาศาล

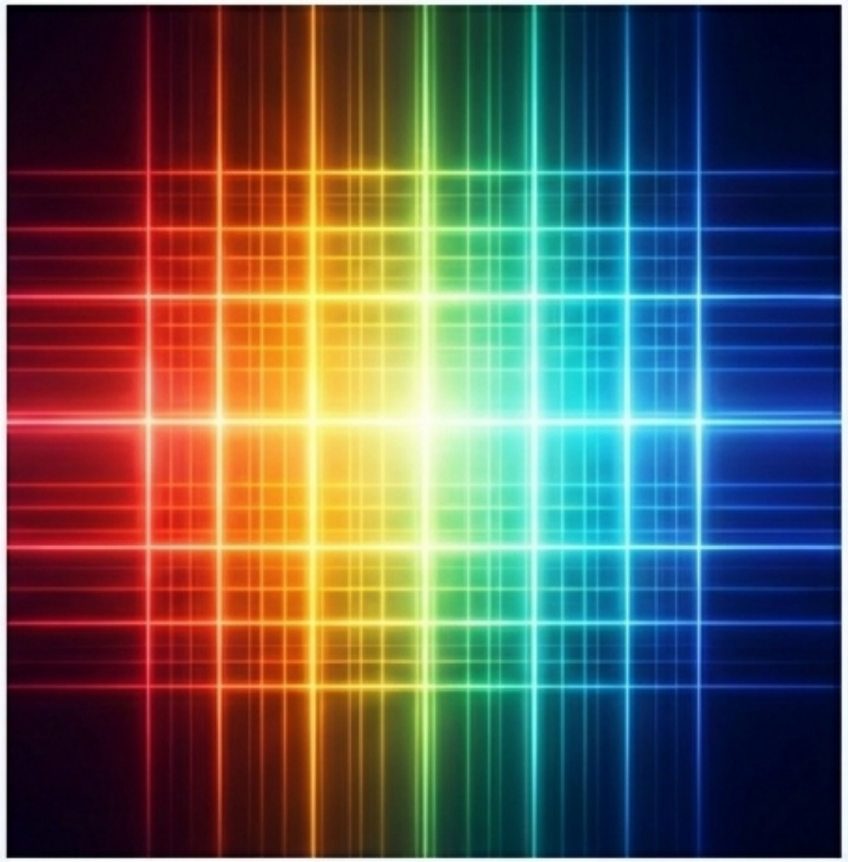
NotebookLM

มาตรฐานใหม่สำหรับ Latent Space

1. **ทลายข้อจำกัด:** ผสานความเข้าใจบริบท (Semantics) และความสมจริง (Pixels) ได้อย่างไร้รอยต่อในโครงสร้างเดียว

2. **ทรงพลังและยืดหยุ่น:** สถาปัตยกรรมอิงความถี่ (Frequency-aware) เปิดประตูสู่การเทรนที่เร็วและประหยัดทรัพยากร

3. **พิสูจน์แล้ว:** ก้าวข้ามขีดจำกัดเดิมด้วยผลลัพธ์ SOTA ในทุกการทดสอบ ทั้งคุณภาพภาพ (FID, PSNR) และประสิทธิภาพ (GFLOPs)



UAE ไม่ใช่แค่โมเดลเข้ารหัส แต่คือเลนส์คู่ใหม่ที่ช่วยให้ AI มองเห็นโลกอย่างที่มันควรจะเป็น

github.com/WeichenFan/UAE

NotebookLM