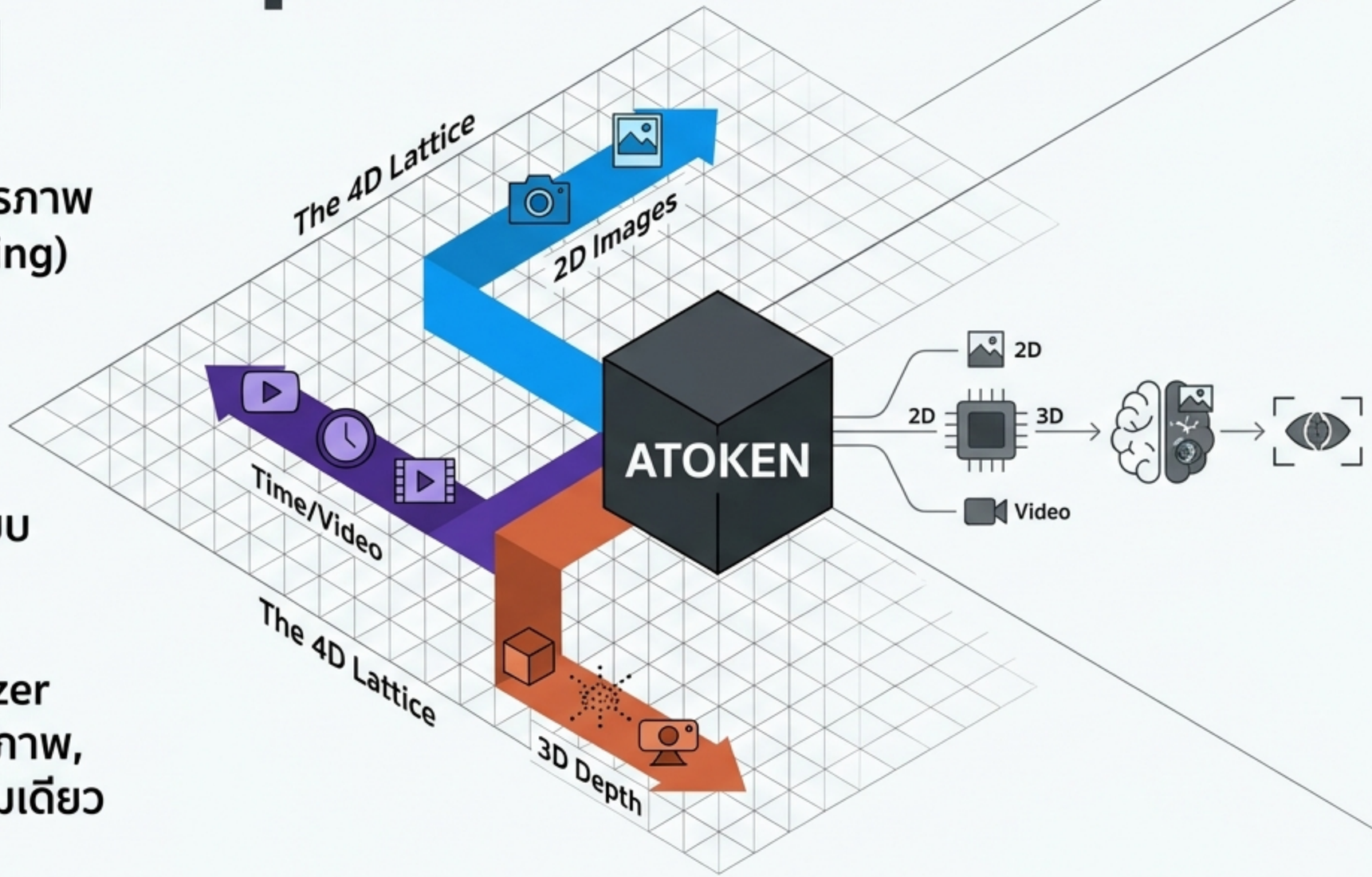
จุดเริ่มต้นของยุคการหลอมรวม Vision AI

โมเดลภาษา (LLMs) ก้าวสู่อิสรภาพด้วย BPE (Byte Pair Encoding) ที่แปลงทุกตัวอักษรเป็นหนึ่งมาตรฐาน

Vision AI ยังคงติดอยู่ในข้อจำกัดของมิติภาพและรูปแบบงานที่แยกส่วนกัน

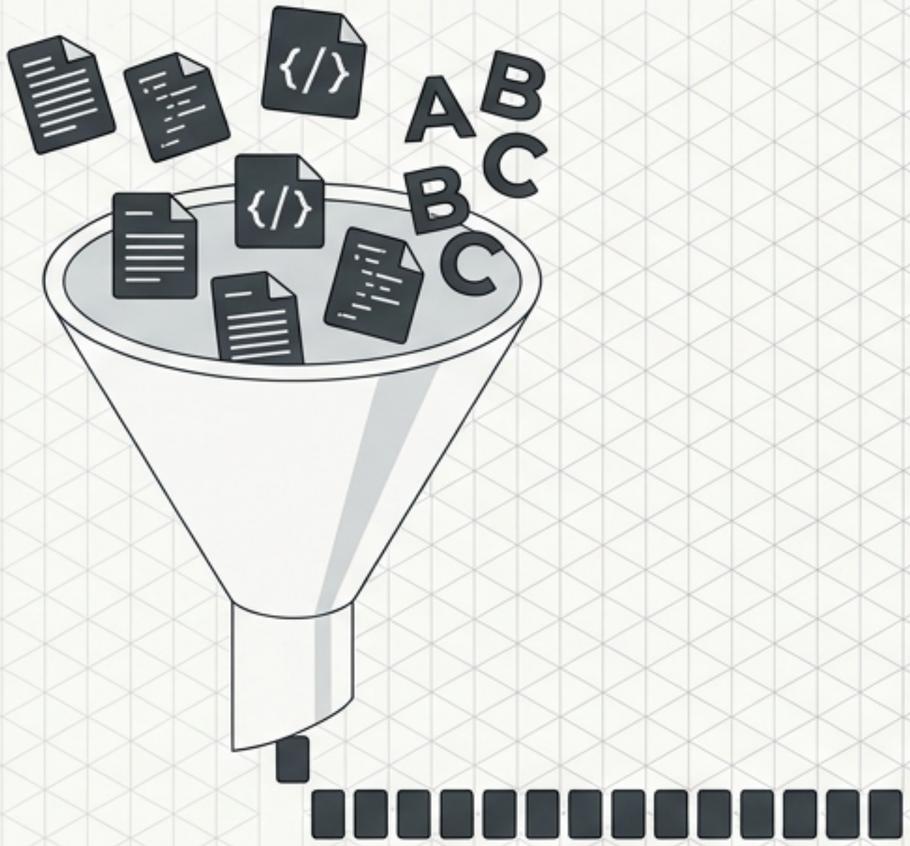
ATOKEN คือ Visual Tokenizer ตัวแรกที่ทลายกำแพงระหว่างภาพ, วิดีโอ และ 3D ในสถาปัตยกรรมเดียว



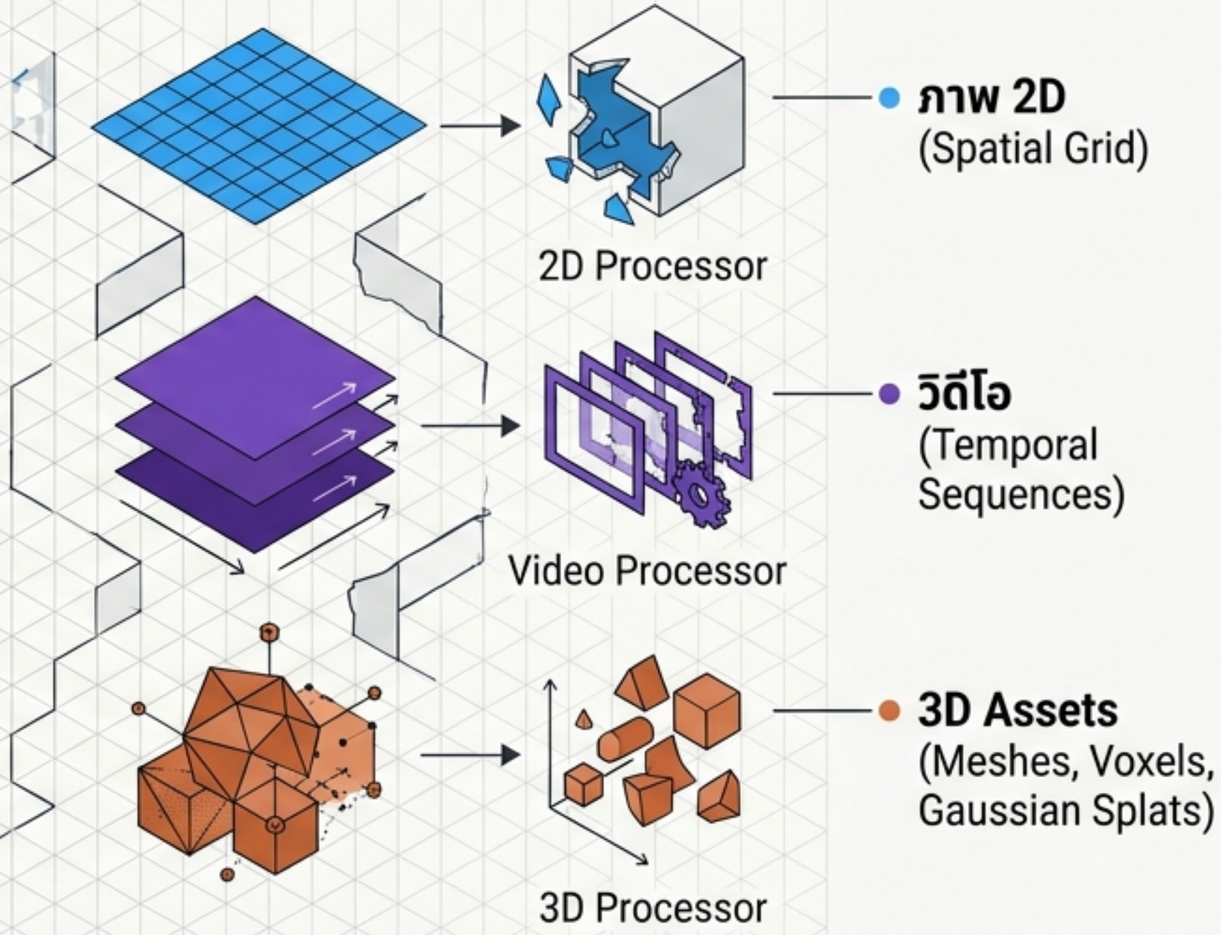
NotebookLM

ปัญหาความแตกแยกในโลกของ Visual Representation

โลกของข้อความ (Text): ประสบความสำเร็จด้วยโทเคนมาตรฐานเดียวที่รองรับทุกข้อมูล



โลกของการมองเห็น (Vision): ถูกจำกัดด้วยความแตกต่างเชิงโครงสร้าง



NotebookLM

ทางแยกที่ต้องเลือก: ความละเอียด หรือ ความเข้าใจ

Reconstruction (SD-VAE, VQGAN)

รักษาความละเอียดระดับพิกเซล (Low-level details)
แต่สูญเสียความหมายเชิงความเข้าใจ (Semantics)

Understanding (CLIP, SigLIP2)

สกัดฟีเจอร์ระดับสูงเพื่อเชื่อมโยงกับข้อความ
แต่ไม่สามารถสร้างภาพย้อนกลับได้ (Cannot reconstruct)

NotebookLM

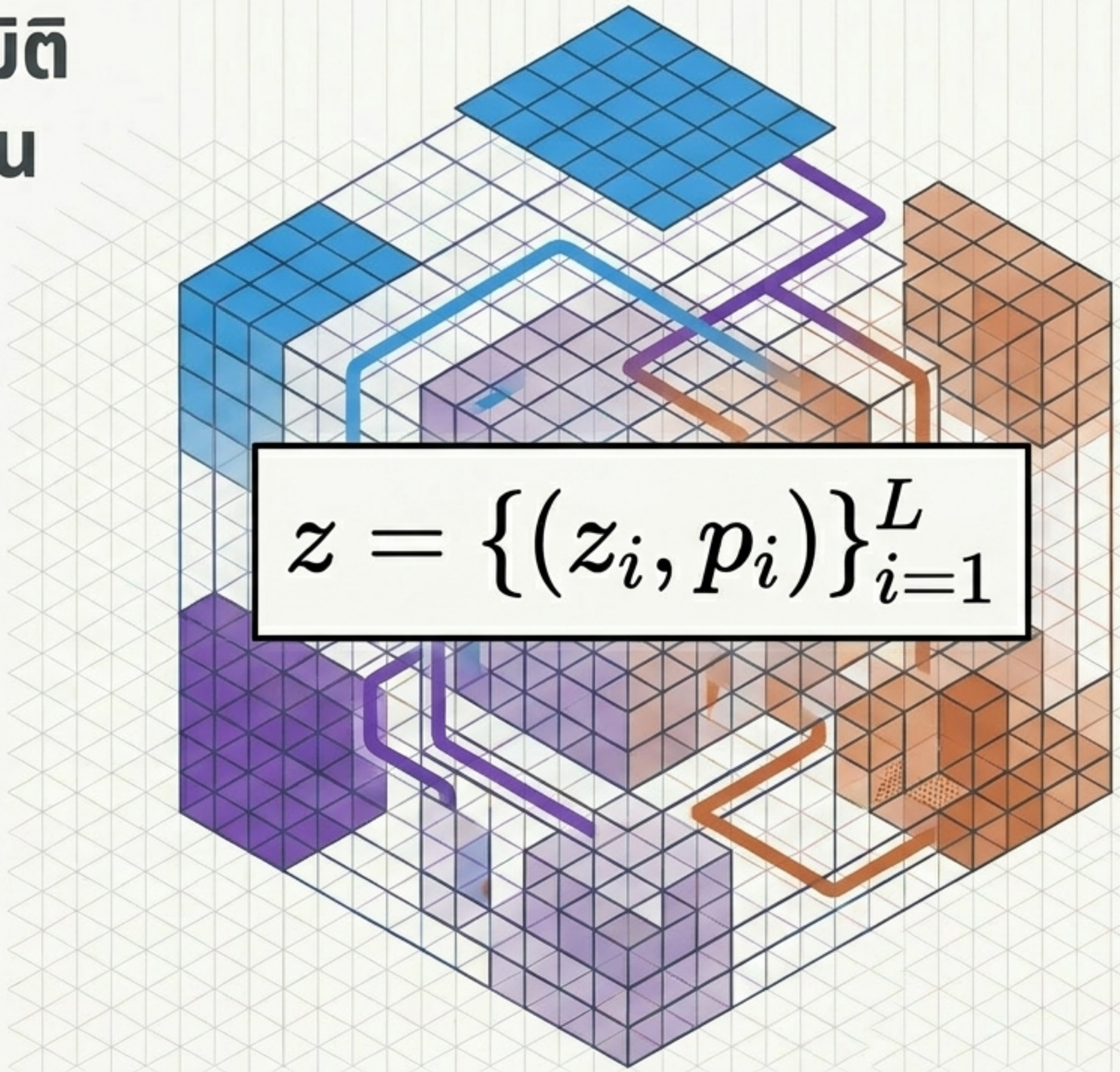
สมการเดียวที่รวมทุกมิติการมองเห็นเข้าด้วยกัน

ขอแนะนำ ATOKEN

โมเดลที่เข้ารหัสภาพ, วิดีโอ และ 3D ลงใน Sparse 4D Latent Space ร่วมกัน

ทลายข้อจำกัดเดิม: ทำได้ทั้งความเข้าใจ (Understanding) และการสร้างภาพ (Generation) ในตัวเดียว

รองรับความละเอียดแบบ Native และ **ความยาวของเวลาอย่างอิสระ**ผ่านสถาปัตยกรรม Pure Transformer



NotebookLM

ATOKEN เป็นโมเดลเดียวที่ครอบคลุมทุก Modality, ทุก Task และรองรับทั้ง Continuous/Discrete Tokens โดยปราศจาก GAN

Models	MODALITY:			TASK:		TOKEN:		FEATURES:	
	2D	Video	3D	Recon	Understand	Discrete	Cont.	GAN-free	Native Res
SD-VAE		✗	✗		✗		✗	✗	
VQGAN		✗	✗		✗		✗	✗	
OmniTokenizer								✗	✗
Trellis-SLAT									
SigLIP2	✗	✗	✗	✗			✗		✗
VILA-U			✗	✗					
ATOKEN	✓	✓	✓	✓	✓	✓	✓	✓	✓

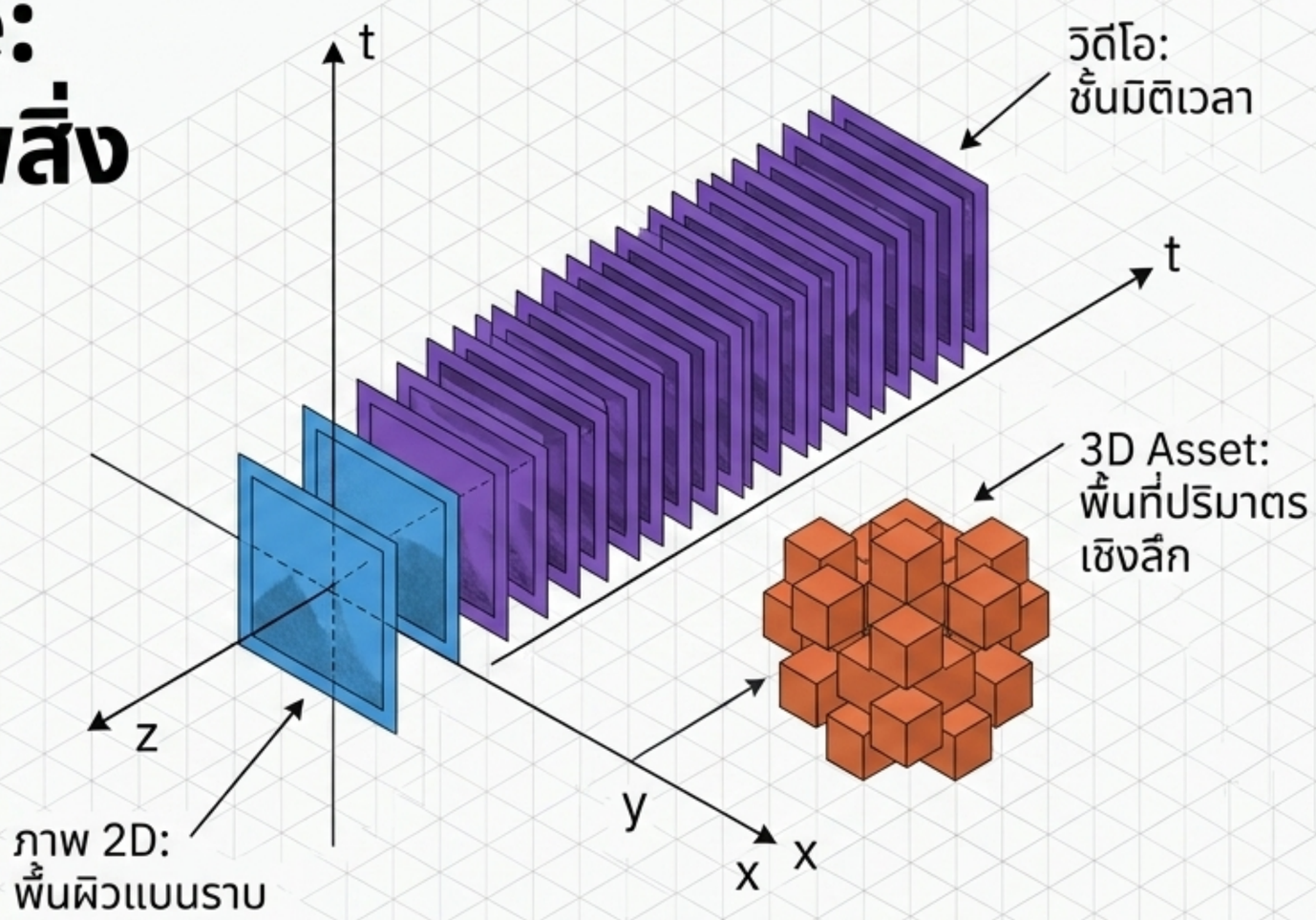
การเปรียบเทียบความสามารถตาม Modality, Task, Token และคุณสมบัติเฉพาะ

NotebookLM

มิติ 4D Latent Space: ที่อยู่อาศัยของทุกสรรพสิ่ง

แทนที่จะแยกโมเดล ATOKEN ใช้ Sparse Representation เพื่อวางข้อมูลแต่ละประเภทลงใน Subspace ของตัวเองอย่างเป็นธรรมชาติ

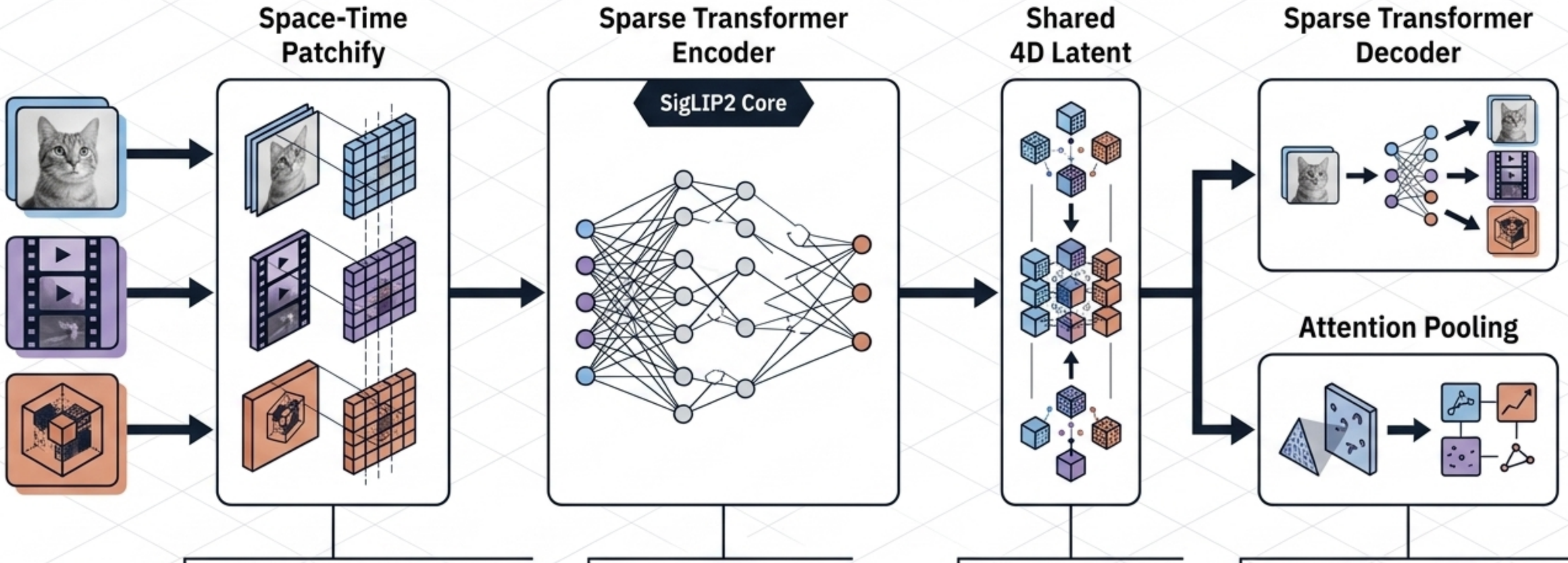
ขับเคลื่อนด้วย 4D Rotary Position Embeddings (RoPE) ที่ทำให้ Transformer เข้าใจตำแหน่งสัมพันธ์ข้ามมิติได้อย่างแม่นยำ



t = เวลา, x,y = พิกัดเชิงพื้นที่, z = ความลึก

NotebookLM

สายพานการผลิตแบบรวมศูนย์ (The Unified Architecture)



1. **Space-Time Patchification:** หั่นข้อมูลทุกมิติเป็นบล็อกเวลาและพื้นที่

2. **4D RoPE Encoder:** ประมวลผลด้วย Sparse Transformer (ต่อยอดจาก SigLIP2)

3. **Shared Latent:** จุดศูนย์รวมข้อมูลที่พร้อมต่อยอด

4. **Task-Specific Outputs:** แยกส่วนการทำงานไปยัง Decoder (Reconstruction) หรือ Attention Pooling (Understanding)

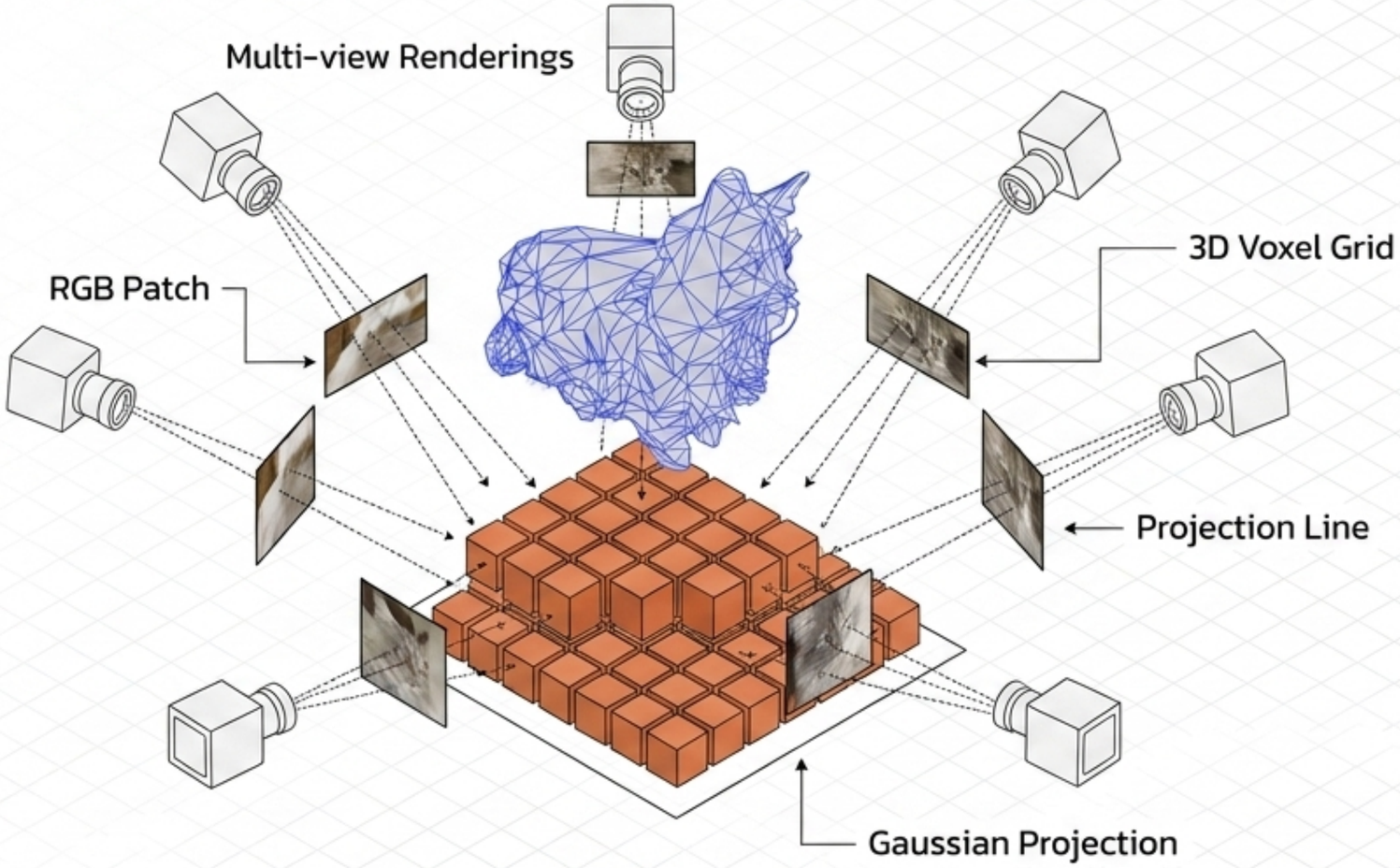
NotebookLM

การเชื่อม 3D เข้าสู่ระบบ Tokenizer

การประมวลผล 3D อาศัยการเรนเดอร์ภาพจากหลายมุมมอง (Multi-view Renderings)

ดึงแพทช์ภาพ (RGB patches) จากมุมกล้องที่ใกล้ที่สุดมาบรรจุลงใน 3D Voxel

ถอดรหัสด้วย Gaussian Splatting เพื่อสร้าง 3D Asset กลับมาได้อย่างสมจริงโดยใช้โครงสร้างเดียวกับ 2D

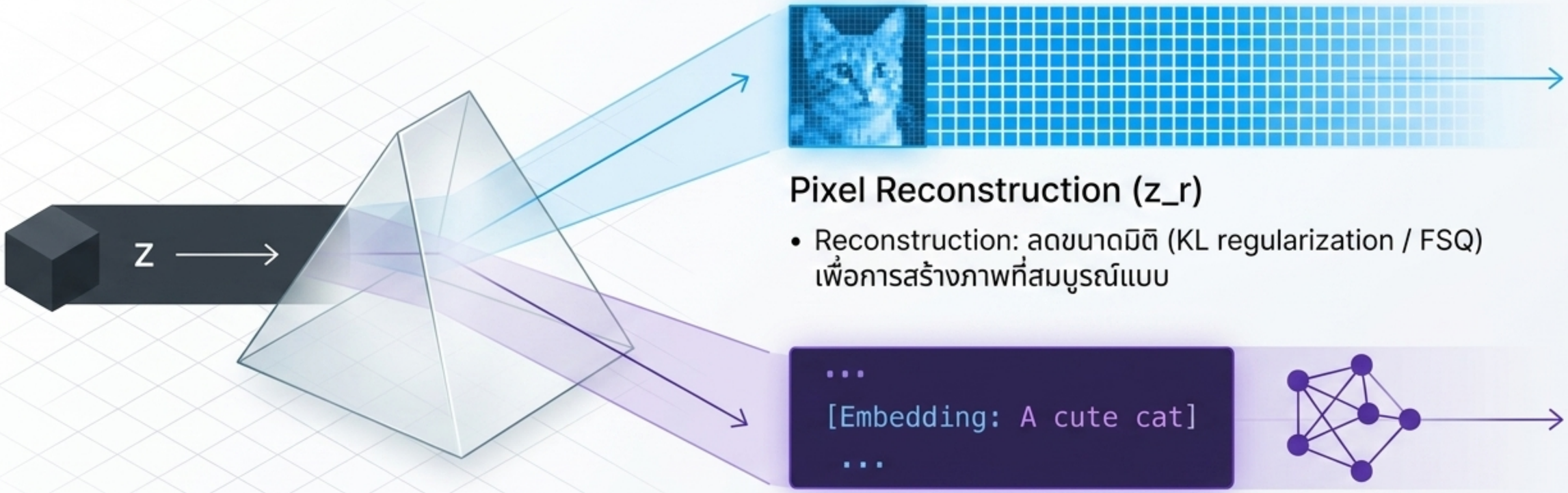


Multi-view → Voxel Mapping → Gaussian Projection

NotebookLM

พลังของ Dual-Head Projection

ฟีเจอร์ที่ถูกเข้ารหัส (z) ถูกฉายออกเป็นสอง**เส้นทางคู่ขนาน**โดยไม่ต้องแยกสถาปัตยกรรม



Pixel Reconstruction (z_r)

- Reconstruction: ลดขนาดมิติ (KL regularization / FSQ) เพื่อการสร้างภาพที่สมบูรณ์แบบ

Semantic Understanding (z_s)

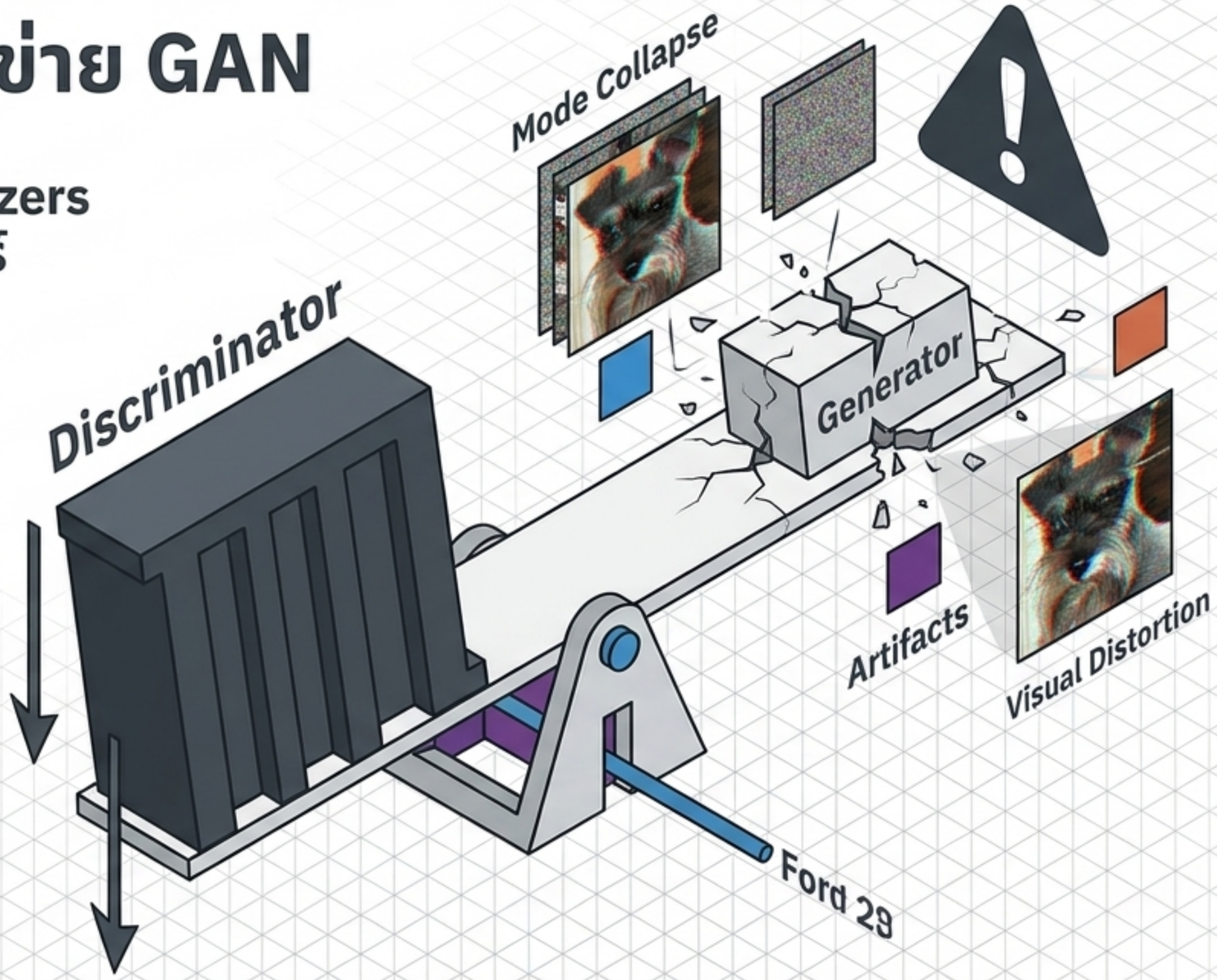
- Understanding: รวมฟีเจอร์ผ่าน Attention Pooling เพื่อจับคู่กับความหมายของข้อความผ่าน SigLIP2

NotebookLM

คอขวดทางเทคนิค: ความล้มเหลวของโครงข่าย GAN

Transformer-based Visual Tokenizers มักล้มเหลวเมื่อขยายขนาดพารามิเตอร์

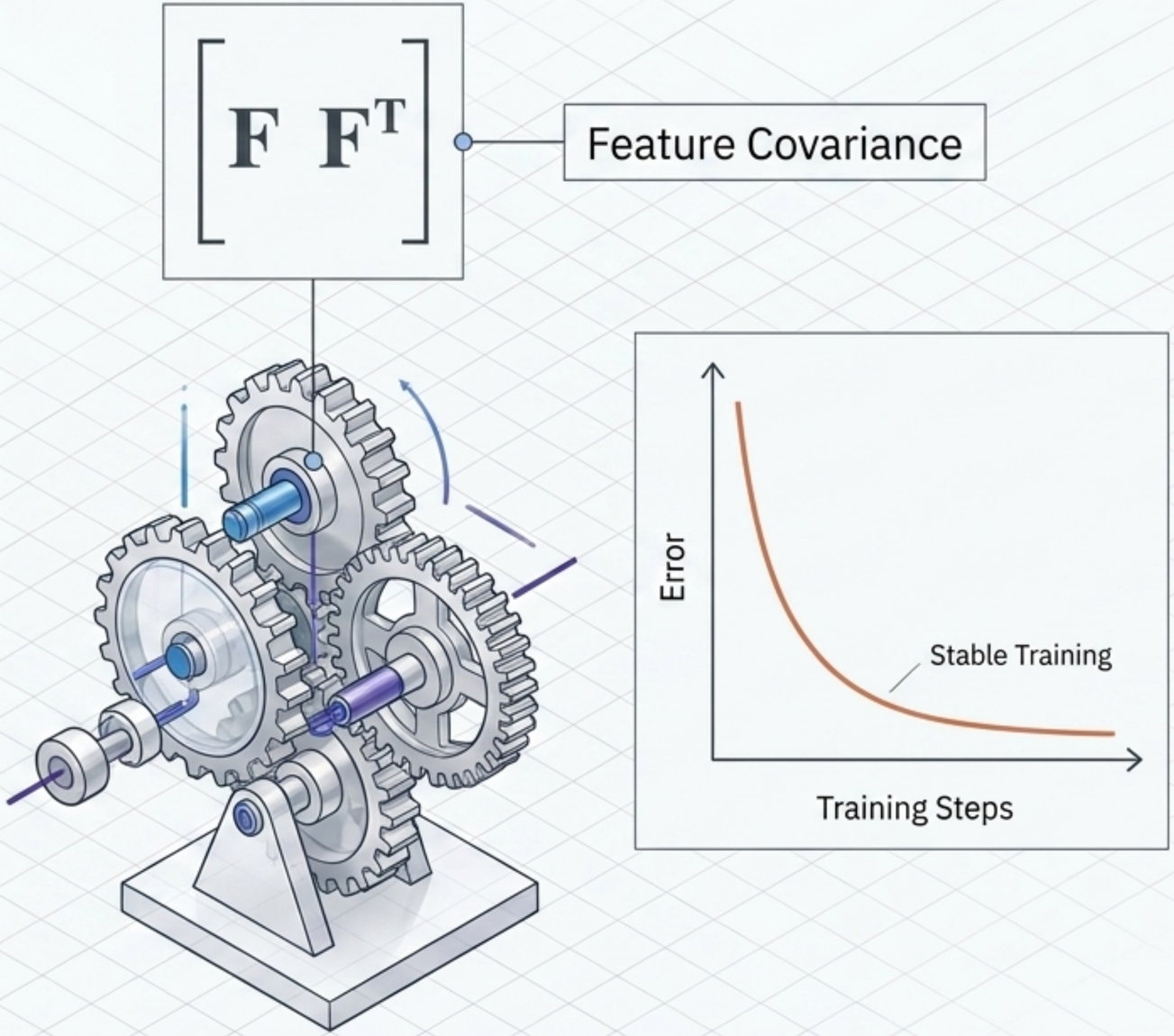
- ต้นเหตุ: โครงข่ายแบบเปรียบปักษ์ (Adversarial Training / GANs) ขาดความเสถียร
- Discriminator มีพลังมากเกินไป กดทับ Generator จนเกิดสภาวะ Mode Collapse และความผิดเพี้ยนของภาพ



NotebookLM

ทางออกใหม่: Adversarial-Free Training ด้วย Gram Loss

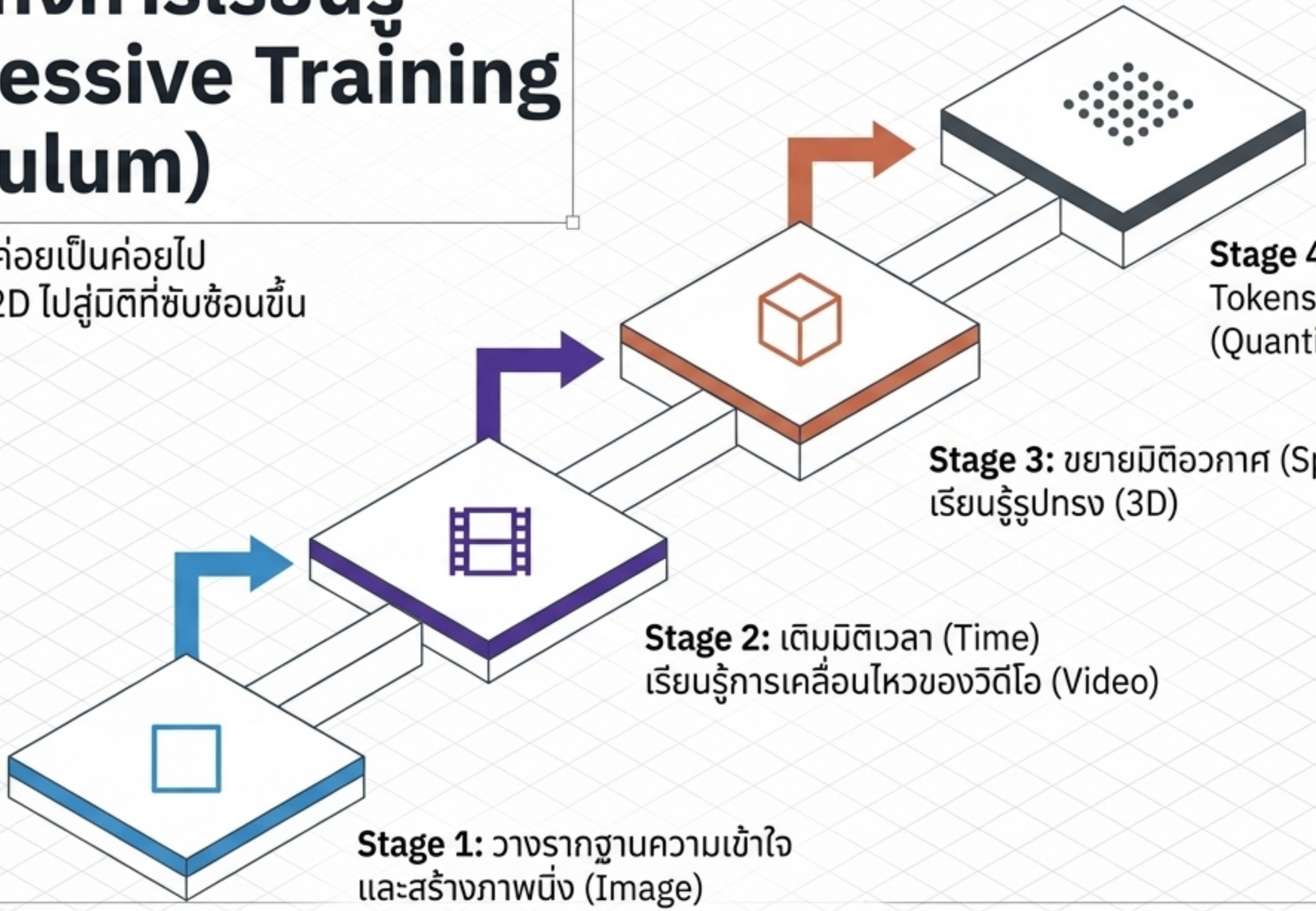
- ATOKEN ตัด GAN ทิ้งอย่างสิ้นเชิง
- อาศัยการรวม Perceptual Loss และ Gram Matrix Loss
- เจาะจงแก้ปัญหาที่ Feature Covariance (ซึ่งส่งผลต่อ Texture และ Style ถึง 86.6% ของค่า Error)
- สร้างภาพคุณภาพสูงสุด (State-of-the-Art rFTD) พร้อมความเสถียรในการฝึกสอนตั้งแต่ต้นจนจบ



NotebookLM

บันไดแห่งการเรียนรู้ (Progressive Training Curriculum)

การฝึกสอนแบบค่อยเป็นค่อยไป
เริ่มจากพื้นฐาน 2D ไปสู่มิติที่ซับซ้อนขึ้น



Stage 1: วางรากฐานความเข้าใจ
และสร้างภาพนิ่ง (Image)

Stage 2: เติมมิติเวลา (Time)
เรียนรู้การเคลื่อนไหวของวิดีโอ (Video)

Stage 3: ขยายมิติอวกาศ (Space)
เรียนรู้รูปทรง (3D)

Stage 4: แปลงเป็น Discrete
Tokens เพื่อเชื่อมต่อกับ LLMs
(Quantization)

NotebookLM

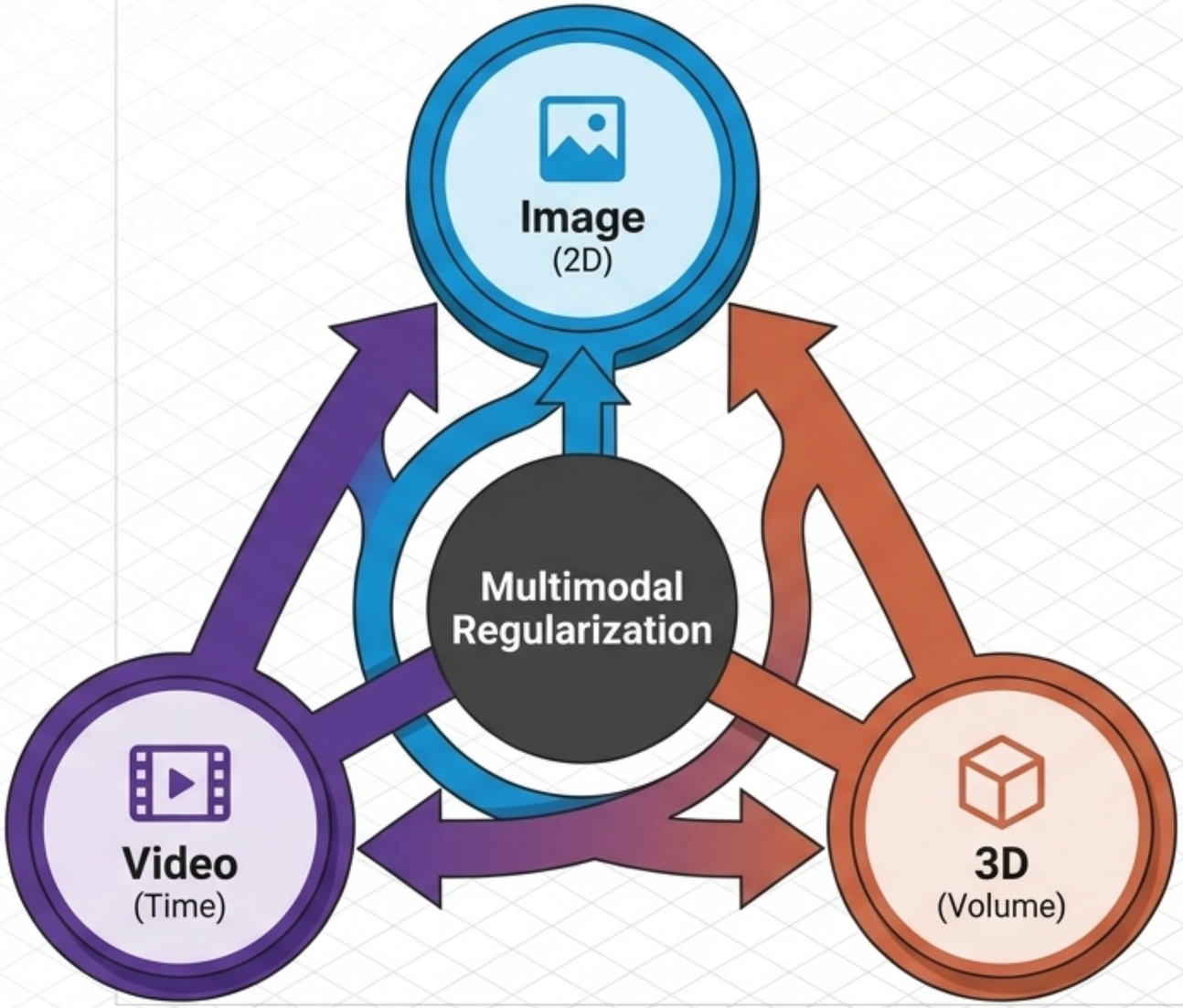
ปรากฏการณ์ Multimodal Regularization (The Aha Moment)

คำถาม: การบังคับให้โมเดลทำได้ทุกอย่าง จะทำให้คุณภาพในแต่ละด้านลดลงหรือไม่?

ความจริงที่ค้นพบ: ตรงกันข้ามอย่างสิ้นเชิง!

การเรียนรู้พลวัตของเวลา (Video) และโครงสร้างเชิงปริมาตร (3D) ช่วยเพิ่มประสิทธิภาพให้กับความสามารถพื้นฐาน 2D

ฟิสิกส์ของการมองเห็นเชื่อมโยงกัน ฐานข้อมูล 4D สร้างรากฐานที่ 'ฉลาดขึ้น' ไม่ใช่แค่ 'กว้างขึ้น'



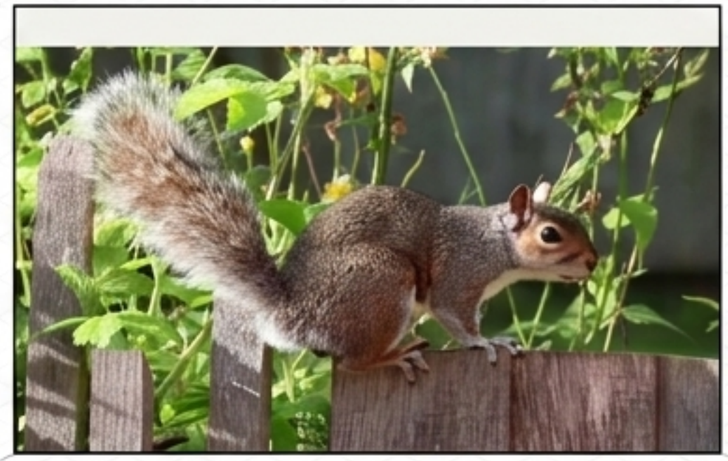
NotebookLM

ข้อพิสูจน์เชิงประจักษ์: ประสิทธิภาพระดับ State-of-the-Art

ความสามารถระดับสูงในทุกมิติ โดยยังคงความเข้าใจเชิงความหมายที่แม่นยำ รองรับตั้งแต่ Continuous Latents ความละเอียดสูง ไปจนถึง Discrete Tokens เพื่อการทำงานร่วมกับ AI ยุคใหม่

(2D Images)

2D Image: 0.21 rFID | 82.2% ImageNet Acc



(Time/Video)

Video: 3.01 rFVD | 40.2% MSRVTT Retrieval

(3D Depth)

3D Asset: 28.28 PSNR | 90.9% 3D Classification

NotebookLM

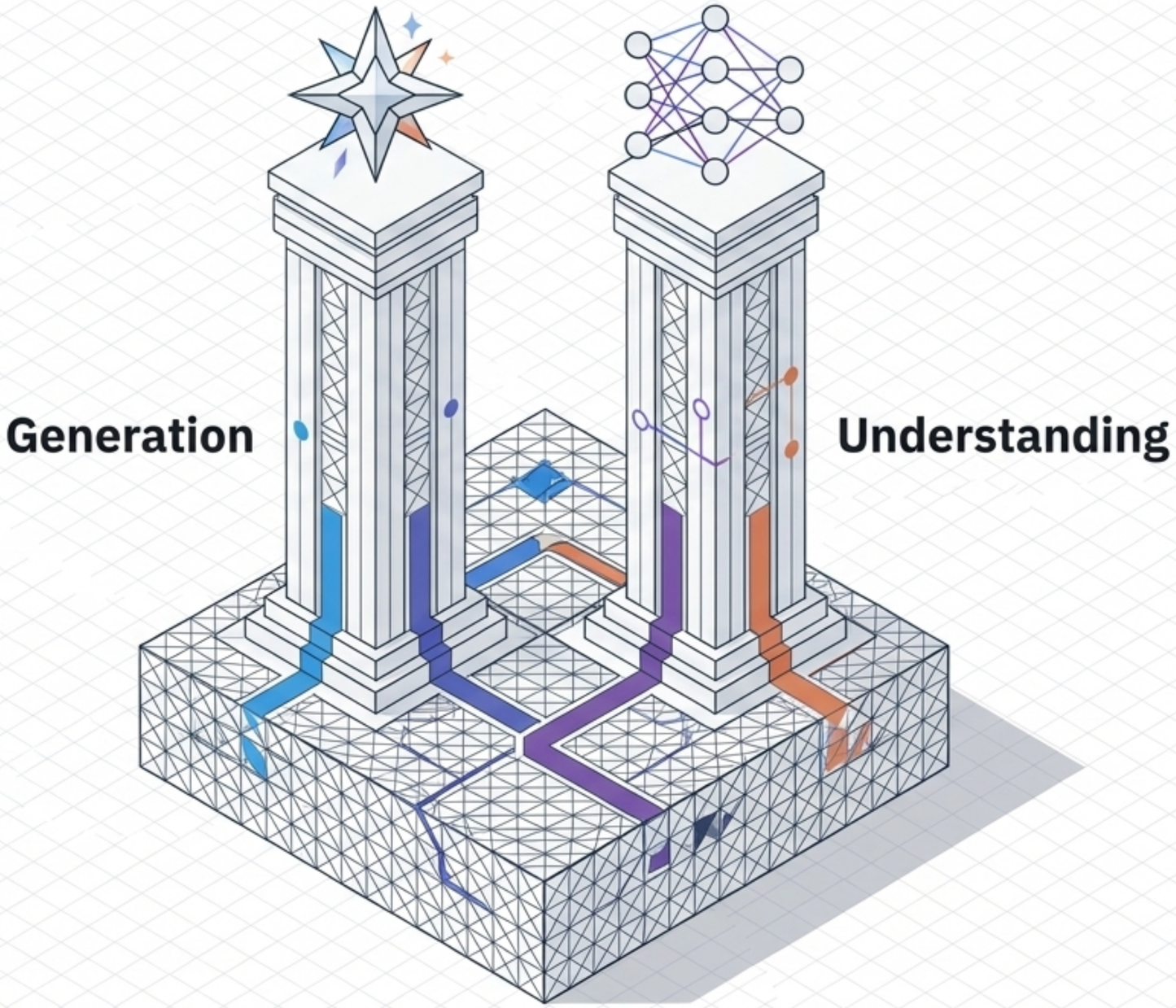
รากฐานใหม่สำหรับ Multimodal AI แห่งอนาคต

ATOKEN ไม่ใช่แค่ Tokenizer แต่คือโครงสร้างพื้นฐาน (Foundation) ของระบบ AI ยุคต่อไป

เข้าใจทุกสิ่ง: ถอดรหัสความหมายของโลกจากทุกมุมมอง

สร้างสรรค์ทุกมิติ: สังเคราะห์ภาพ วิดีโอ และ 3D จากแกนกลางเดียวกัน

หมดยุคของการแยกส่วน (Fragmentation) สู่ยุคทองของการหลอมรวมระบบ AI อย่างแท้จริง (The Grand Unification)



NotebookLM