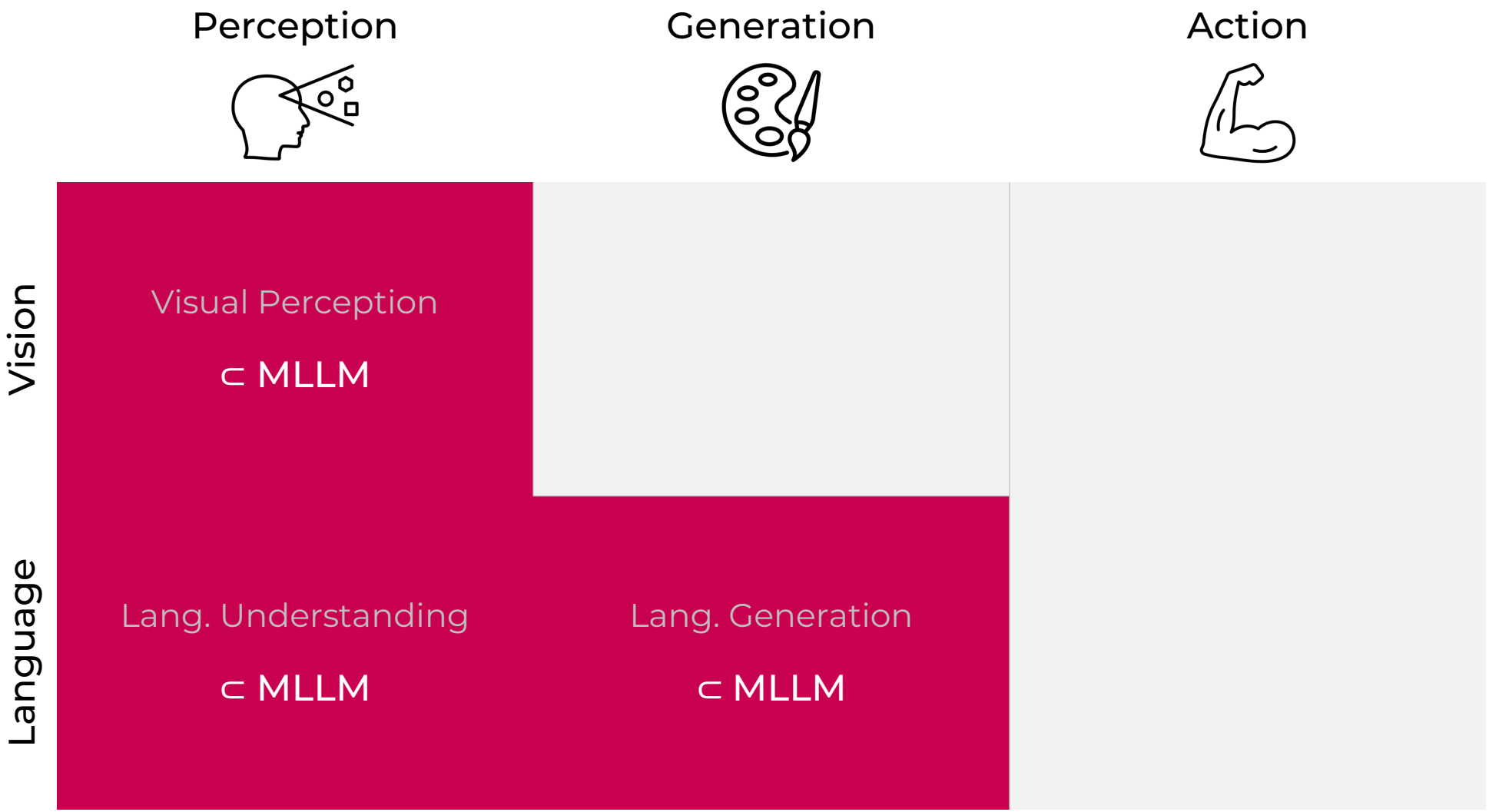
IMEN574/IDSC574 Programming for Data Science

21. **Multimodal Models III**

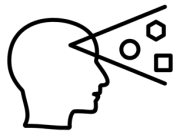
Spring 2026

Instructor: **Sangwoo Mo**

Recap: Multimodal LLM



Visual Generation

	Perception 	Generation 	Action 
Vision	Visual Perception ⊂ MLLM		
Language	Lang. Understanding ⊂ MLLM	Lang. Generation ⊂ MLLM	

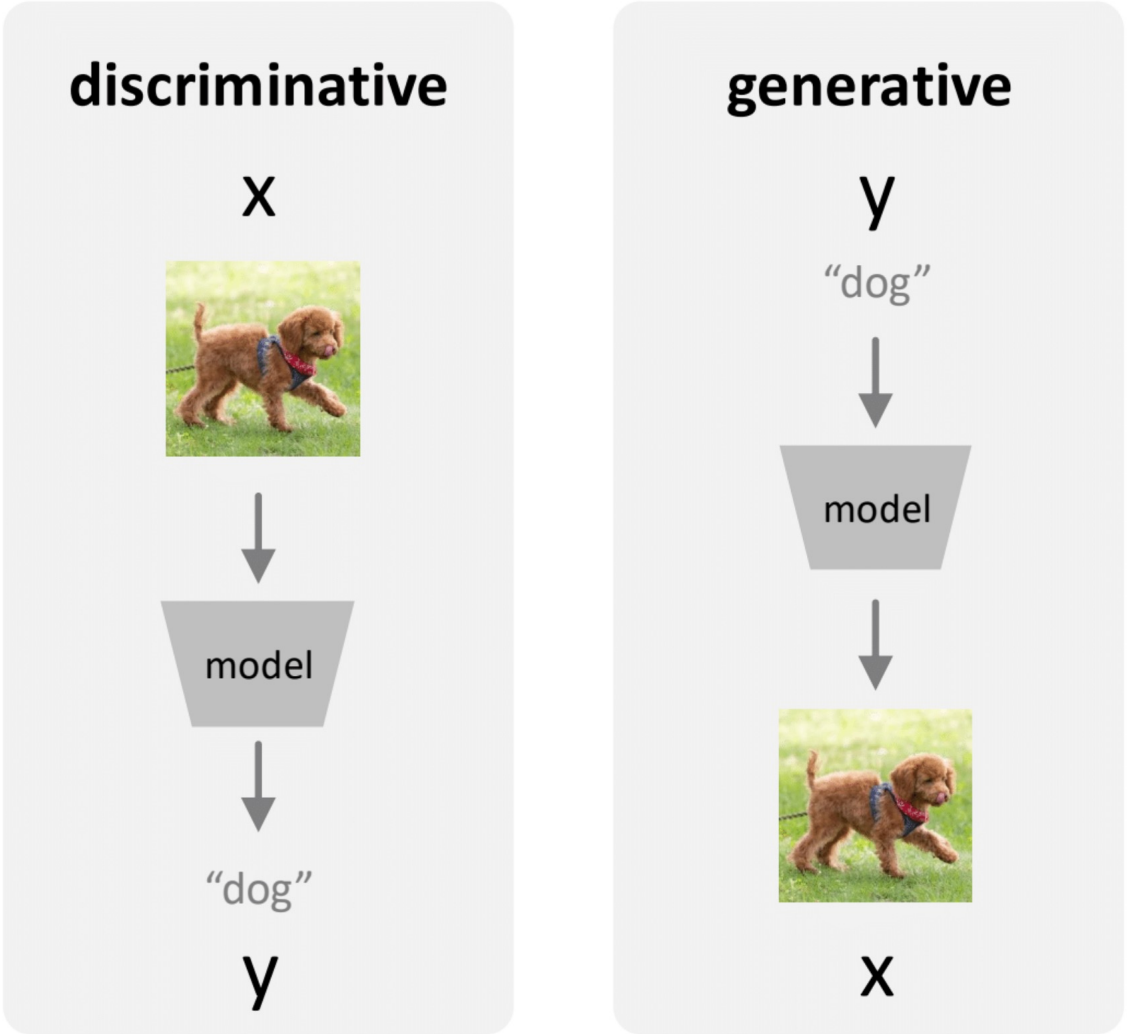
What are Generative Models?

Discriminative Model

- "sample" x ⇨ "label" y
- one desired output

Generative Model

- "label" y ⇨ "sample" x
- many possible outputs

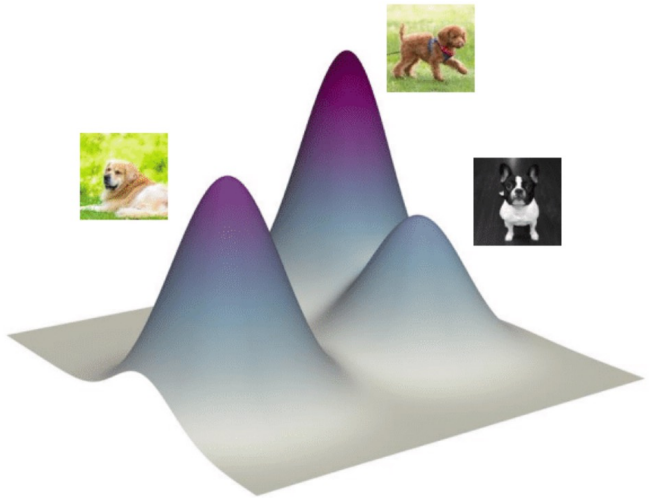


What are Generative Models?

data

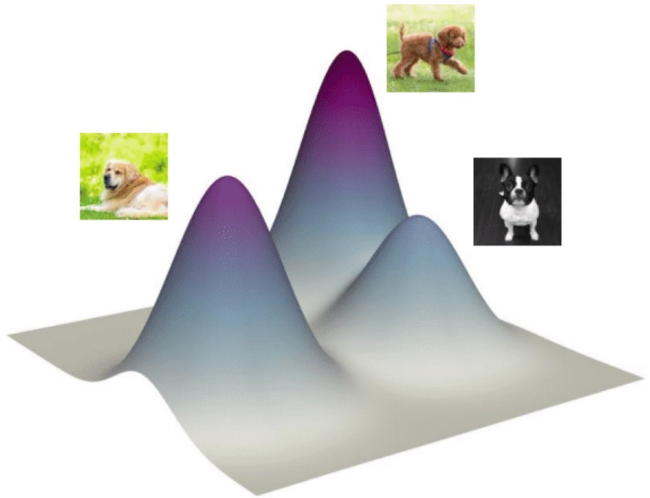


data distribution

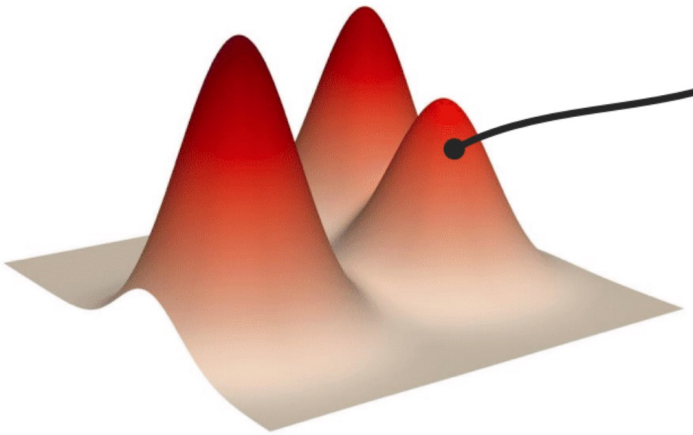


What are Generative Models?

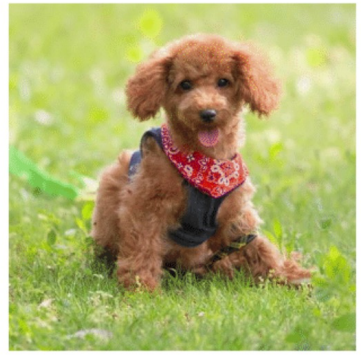
data



data distribution



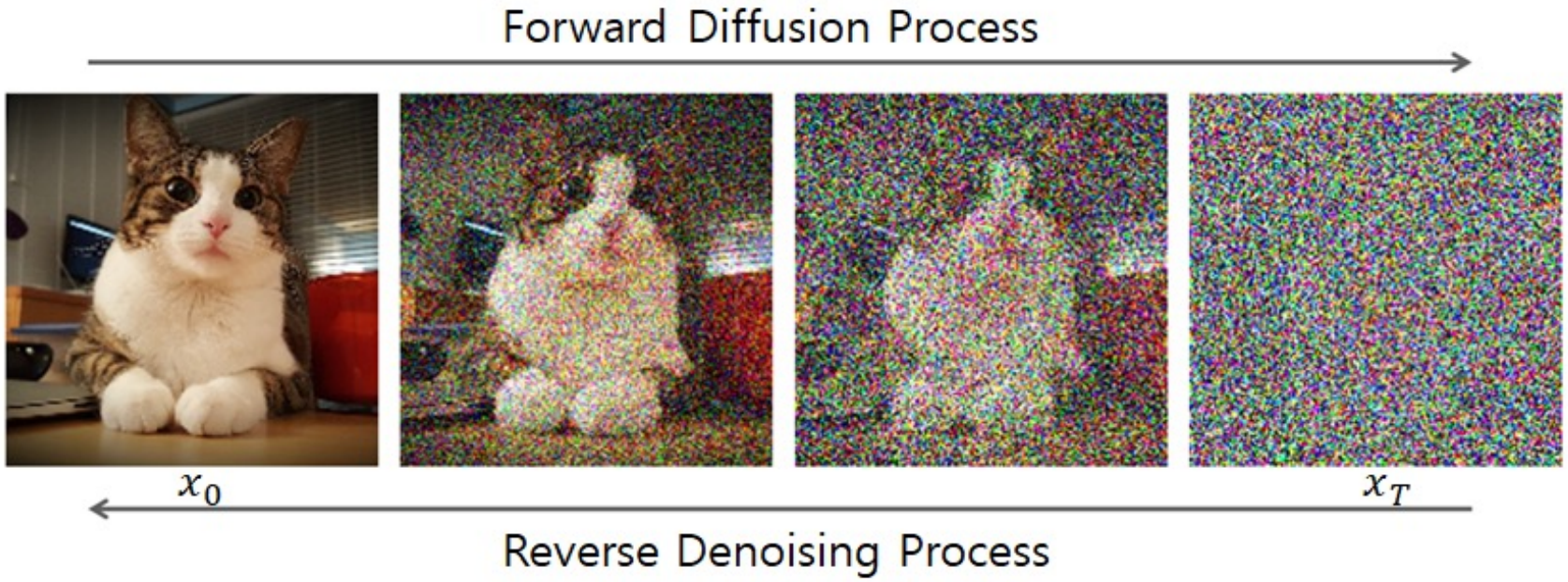
estimated distribution



sample new data

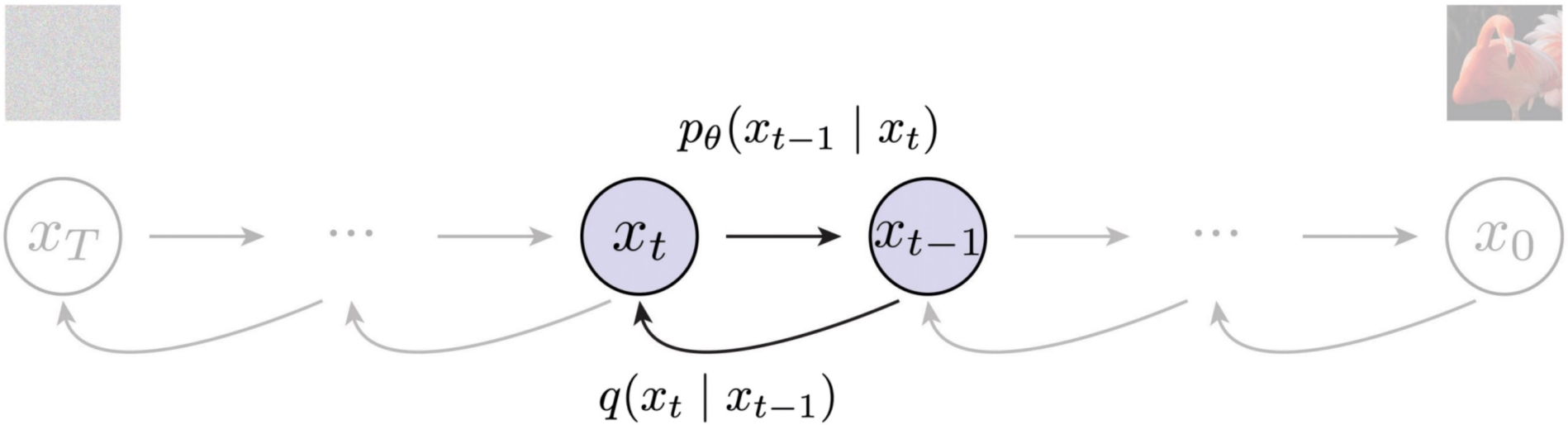
$$x' \sim p(x|y)$$

Current Standard: Diffusion / Flow Matching

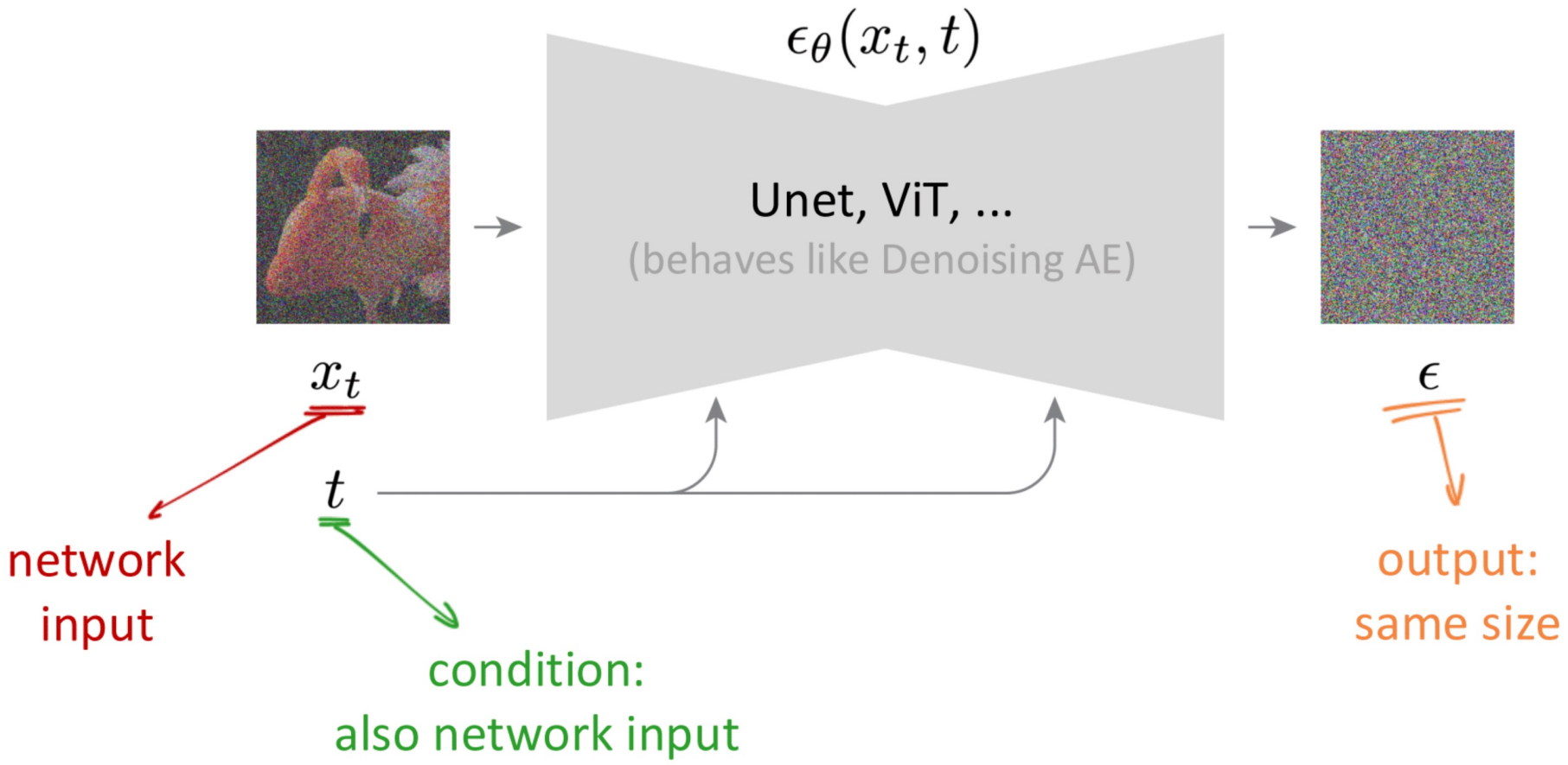


Esser et al. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. arXiv 2024.

Diffusion Models

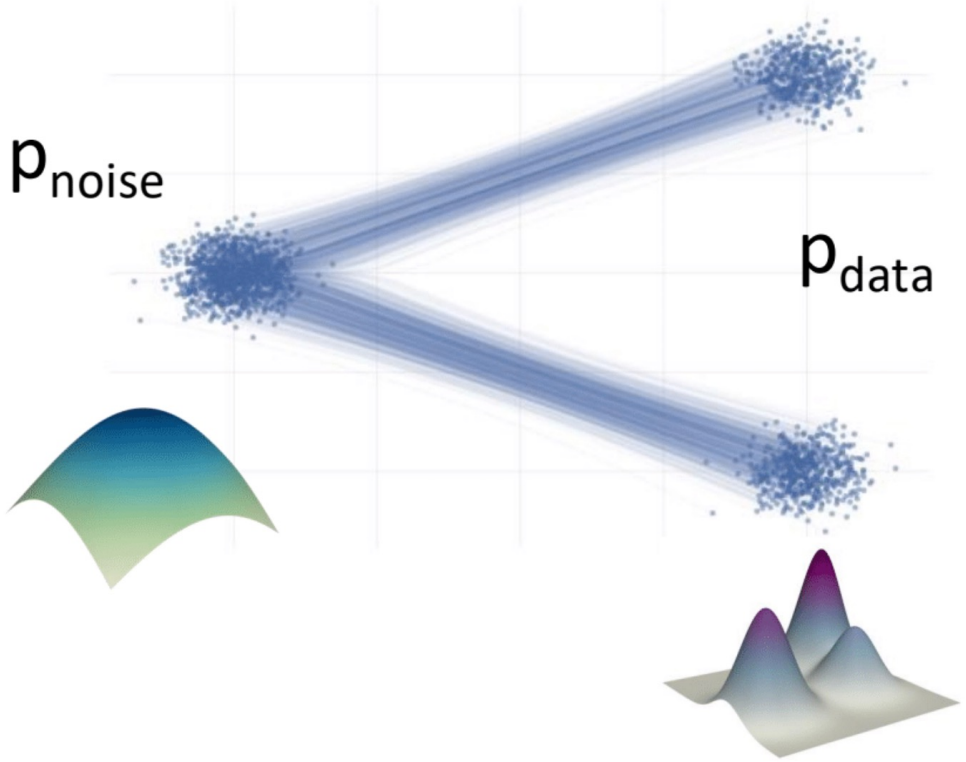


Diffusion Models

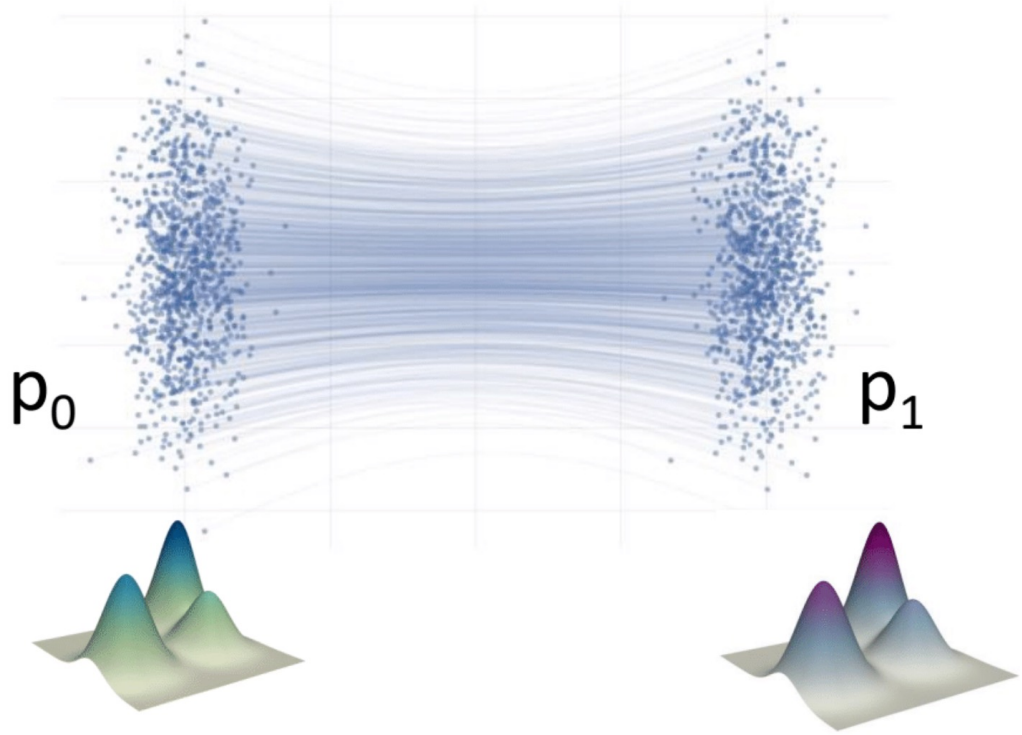


Flow Matching

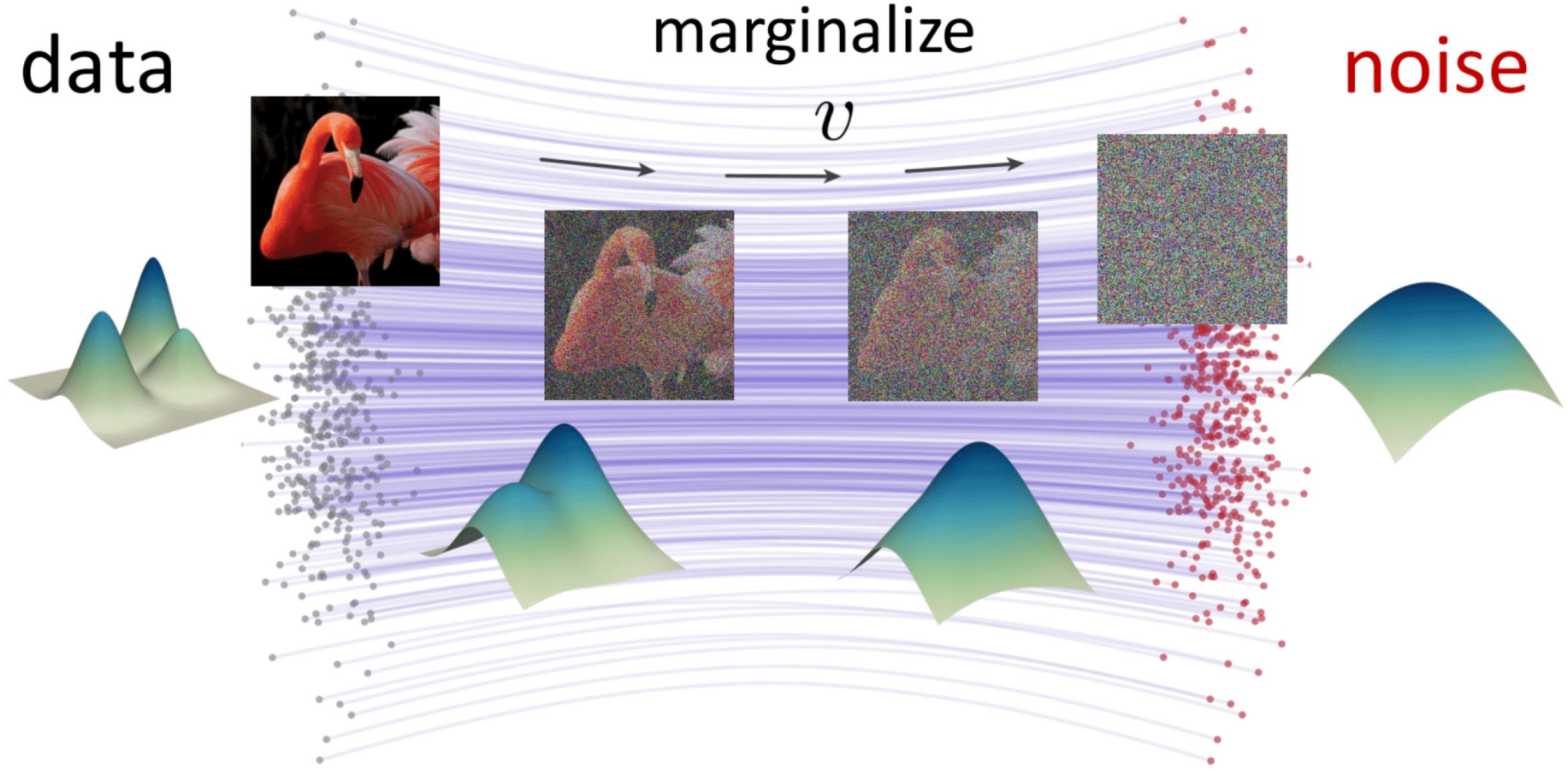
- Noise to Data



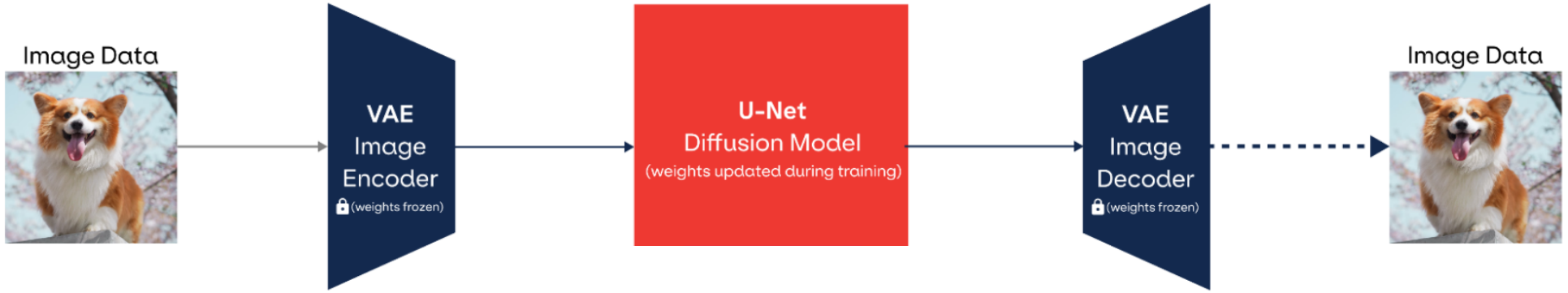
- Conceptually:
  Any to Any



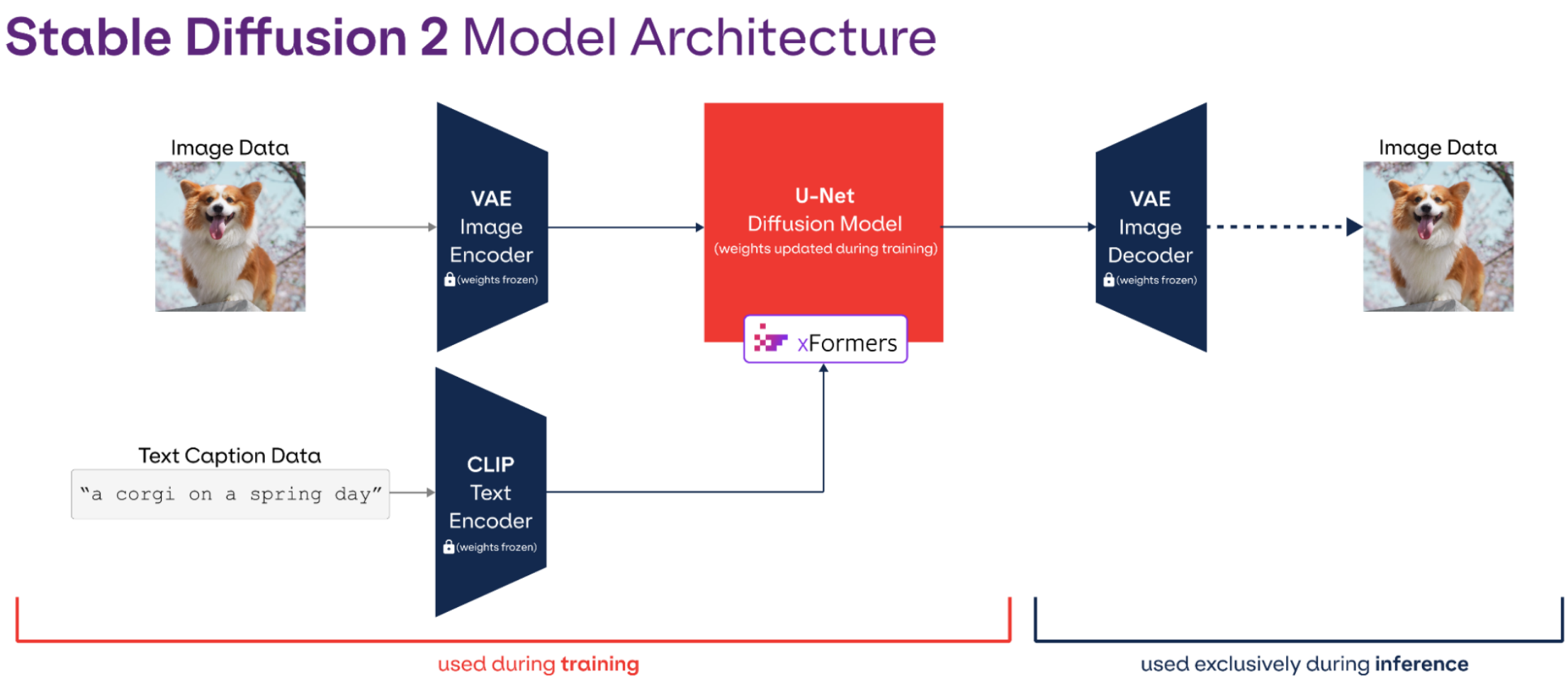
Flow Matching



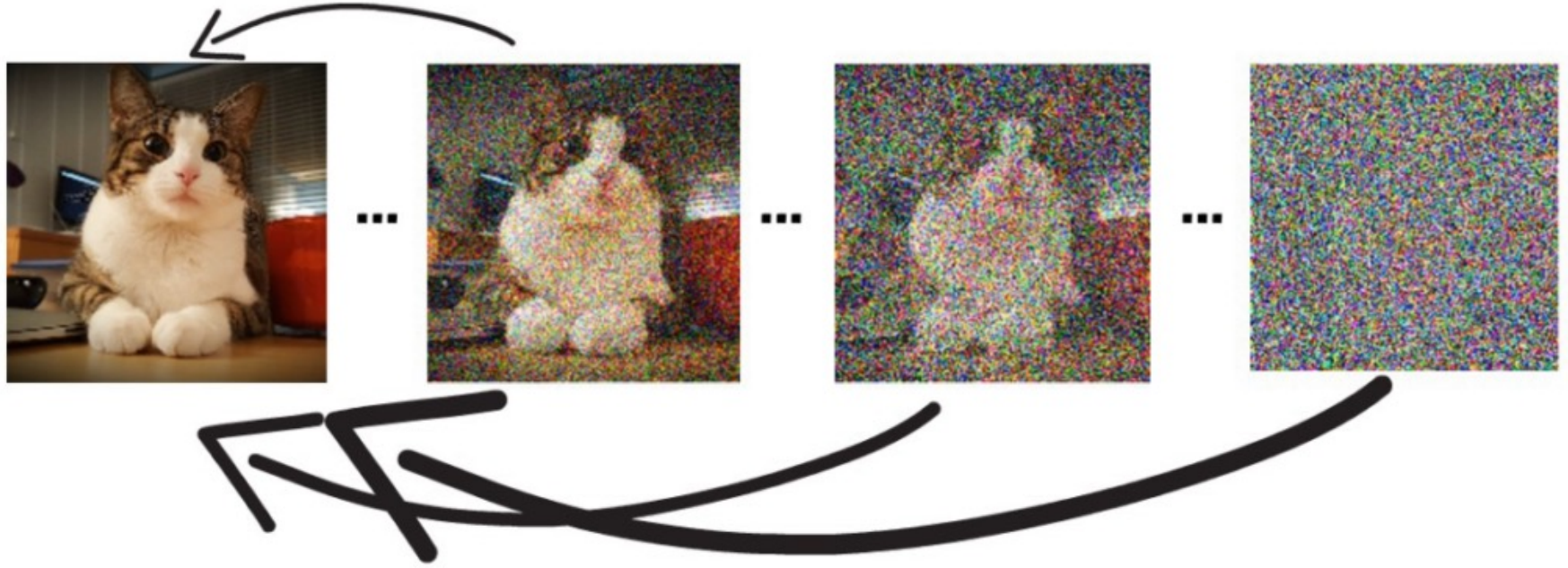
Text-to-Image Generation



Text-to-Image Generation

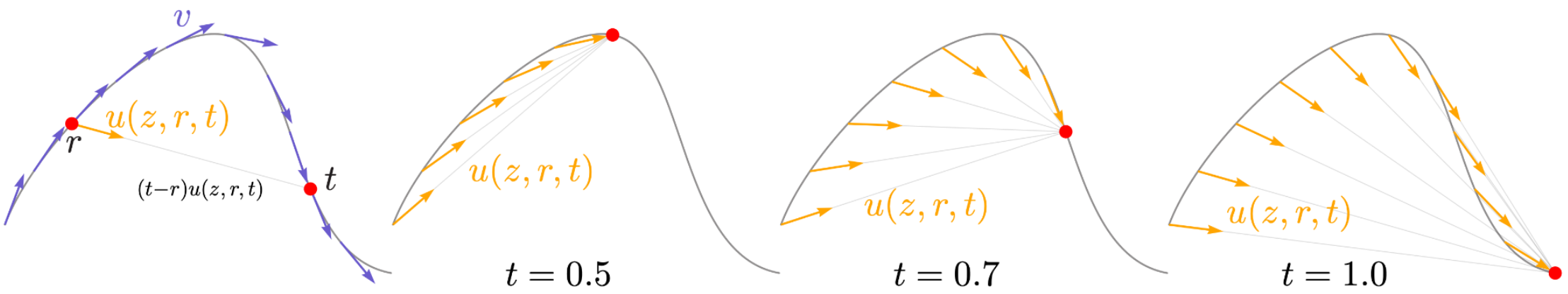


... + 1-Step Distillation



Geng et al. Mean Flows for One-step Generative Modeling. NeurIPS 2025.

… + 1-Step Distillation



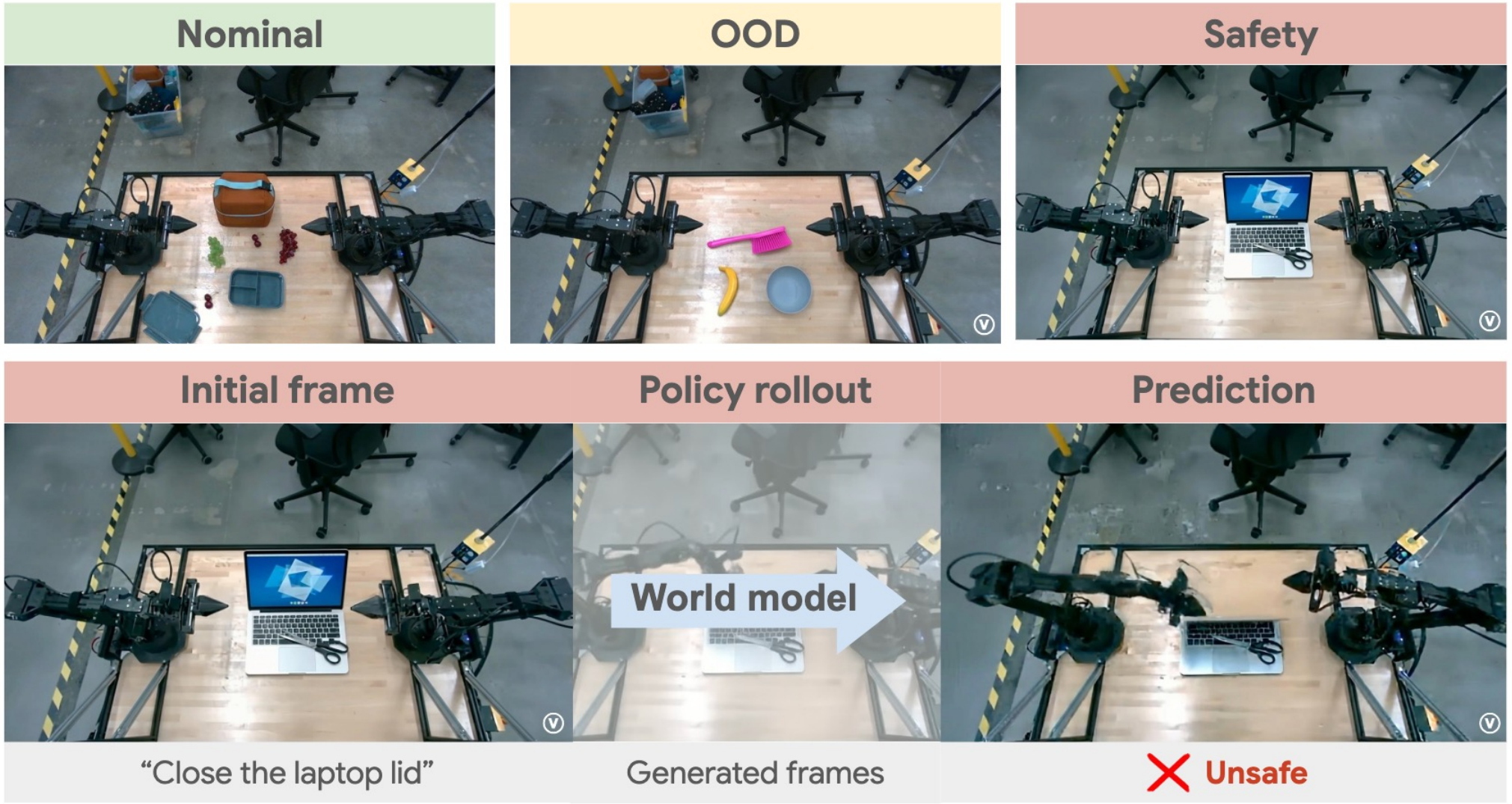
Pretty Image / Video / 3D Generation



More Than... Pretty Image / Video / 3D Generation

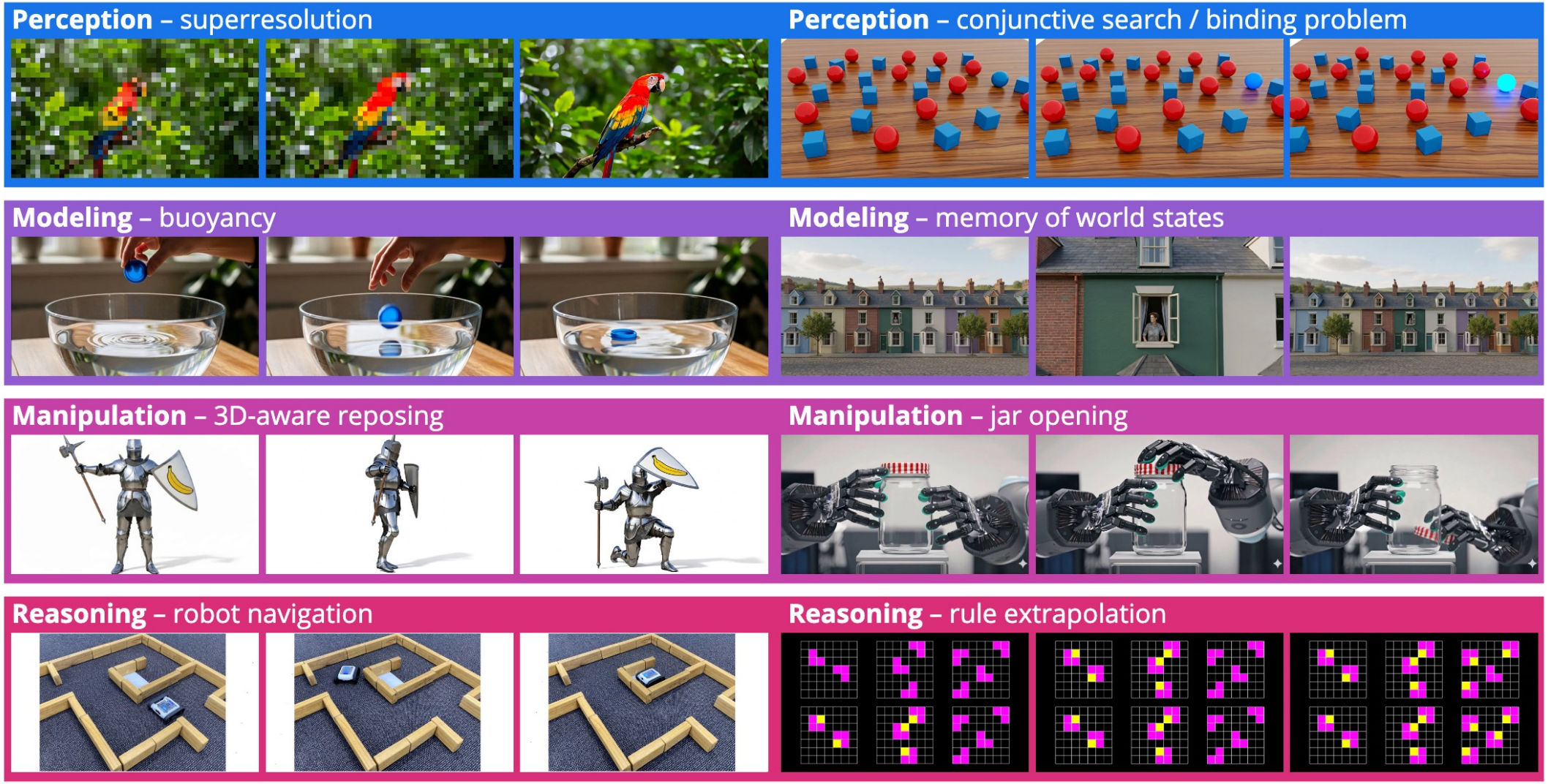


1. Physics Emerges from Visual Generation



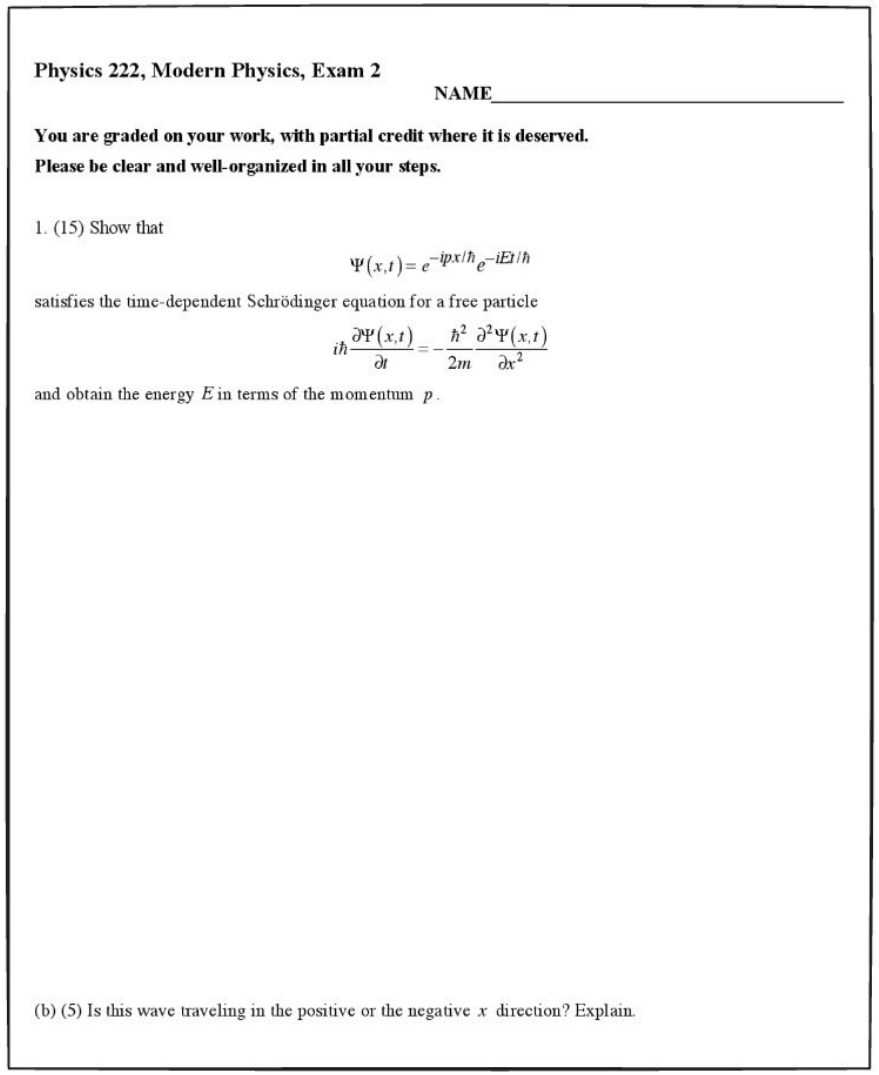
Devin et al. Evaluating Gemini Robotics Policies in a Veo World Simulator. arXiv 2025.

2. ICL Emerges from Visual Generation



Wiedemer et al. Video models are zero-shot learners and reasoners. arXiv 2025.

3. Reasoning Emerges from Visual Generation



Physics 222, Modern Physics, Exam 2

NAME_____

You are graded on your work, with partial credit where it is deserved.
Please be clear and well-organized in all your steps.

1. (15) Show that

$$\Psi(x,t) = e^{-ipx/\hbar}e^{-iEt/\hbar}$$

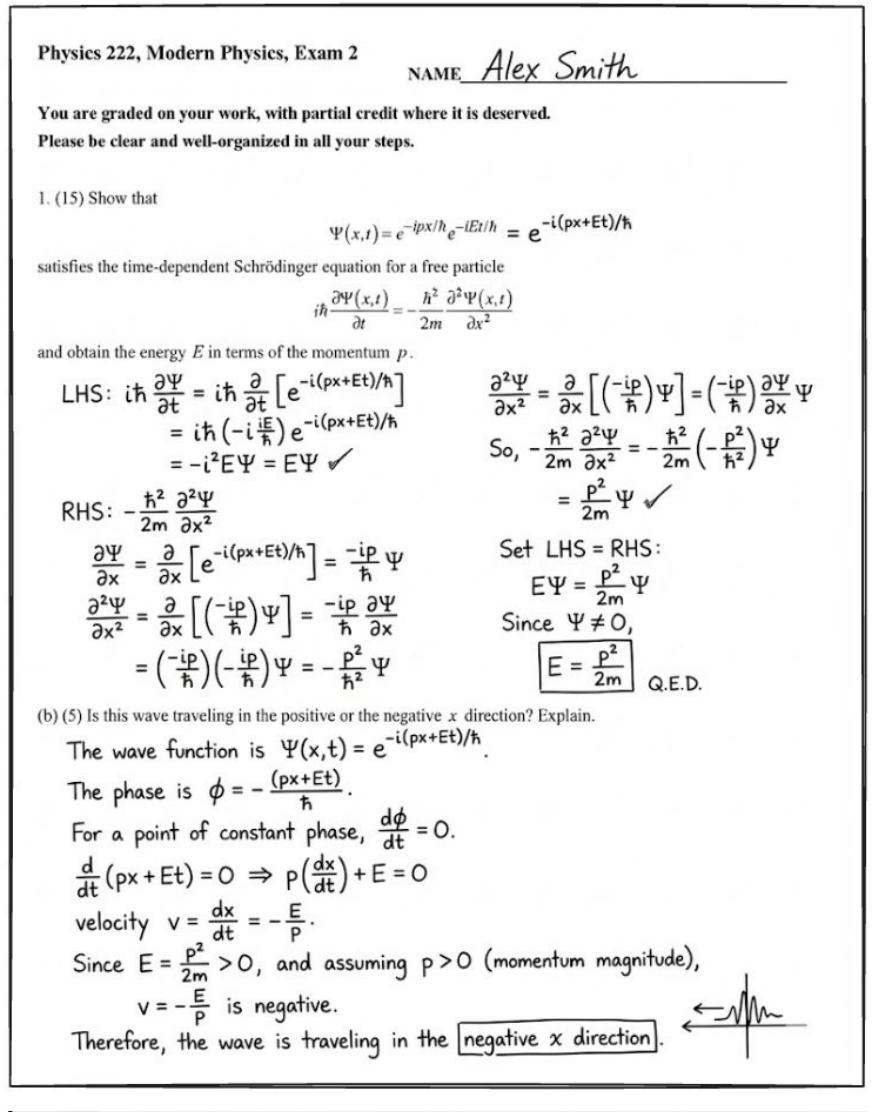
satisfies the time-dependent Schrödinger equation for a free particle

$$i\hbar\frac{\partial\Psi(x,t)}{\partial t} = -\frac{\hbar^2}{2m}\frac{\partial^2\Psi(x,t)}{\partial x^2}$$

and obtain the energy E in terms of the momentum p.

(b) (5) Is this wave traveling in the positive or the negative x direction? Explain.

Gemini Nano Banana Pro



Physics 222, Modern Physics, Exam 2 NAME Alex Smith

You are graded on your work, with partial credit where it is deserved.
Please be clear and well-organized in all your steps.

1. (15) Show that

$$\Psi(x,t) = e^{-ipx/\hbar}e^{-iEt/\hbar} = e^{-i(px+Et)/\hbar}$$

satisfies the time-dependent Schrödinger equation for a free particle

$$i\hbar\frac{\partial\Psi(x,t)}{\partial t} = -\frac{\hbar^2}{2m}\frac{\partial^2\Psi(x,t)}{\partial x^2}$$

and obtain the energy E in terms of the momentum p.

LHS: $i\hbar\frac{\partial\Psi}{\partial t} = i\hbar\frac{\partial}{\partial t}\left[e^{-i(px+Et)/\hbar}\right]$
$= i\hbar\left(-i\frac{iE}{\hbar}\right)e^{-i(px+Et)/\hbar}$
$= -i^2E\Psi = E\Psi$ ✓

RHS: $-\frac{\hbar^2}{2m}\frac{\partial^2\Psi}{\partial x^2}$

$\frac{\partial\Psi}{\partial x} = \frac{\partial}{\partial x}\left[e^{-i(px+Et)/\hbar}\right] = \frac{-ip}{\hbar}\Psi$

$\frac{\partial^2\Psi}{\partial x^2} = \frac{\partial}{\partial x}\left[\left(\frac{-ip}{\hbar}\right)\Psi\right] = \frac{-ip}{\hbar}\frac{\partial\Psi}{\partial x}$
$= \left(\frac{-ip}{\hbar}\right)\left(\frac{-ip}{\hbar}\right)\Psi = -\frac{p^2}{\hbar^2}\Psi$

$\frac{\partial^2\Psi}{\partial x^2} = \frac{\partial}{\partial x}\left[\left(\frac{-ip}{\hbar}\right)\Psi\right] = \left(\frac{-ip}{\hbar}\right)\frac{\partial\Psi}{\partial x}\Psi$

So, $-\frac{\hbar^2}{2m}\frac{\partial^2\Psi}{\partial x^2} = -\frac{\hbar^2}{2m}\left(-\frac{p^2}{\hbar^2}\right)\Psi$
$= \frac{p^2}{2m}\Psi$ ✓

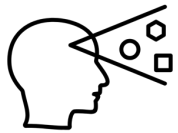
Set LHS = RHS:
$E\Psi = \frac{p^2}{2m}\Psi$
Since $\Psi \neq 0$,
$\boxed{E = \frac{p^2}{2m}}$ Q.E.D.

(b) (5) Is this wave traveling in the positive or the negative x direction? Explain.

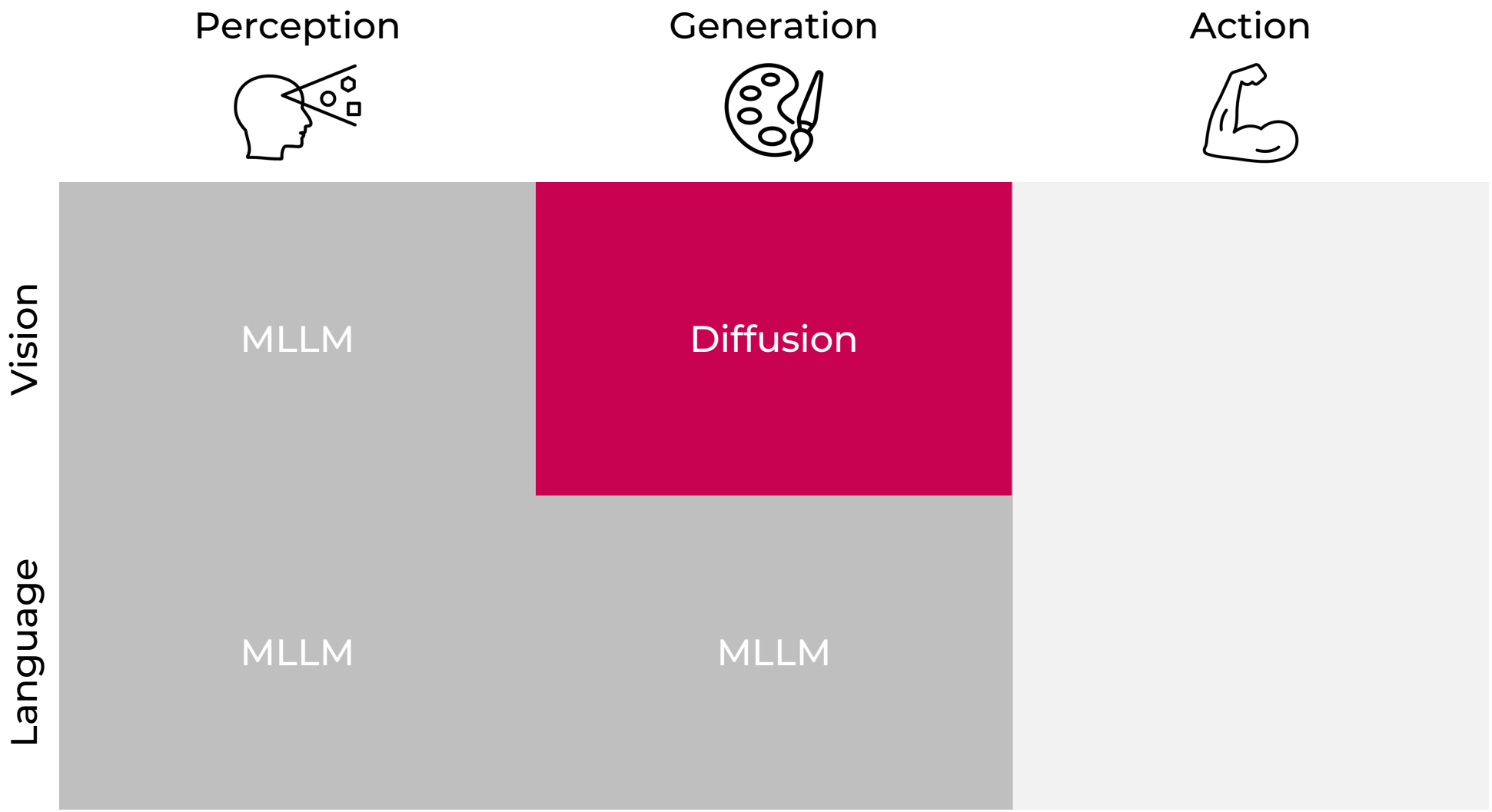
The wave function is $\Psi(x,t) = e^{-i(px+Et)/\hbar}$.
The phase is $\phi = -\frac{(px+Et)}{\hbar}$.
For a point of constant phase, $\frac{d\phi}{dt} = 0$.
$\frac{d}{dt}(px+Et) = 0 \Rightarrow p\left(\frac{dx}{dt}\right) + E = 0$
velocity $v = \frac{dx}{dt} = -\frac{E}{p}$.
Since $E = \frac{p^2}{2m} > 0$, and assuming $p > 0$ (momentum magnitude),
$v = -\frac{E}{p}$ is negative.
Therefore, the wave is traveling in the negative x direction.

https://x.com/karpathy/status/1992655330002817095/photo/1

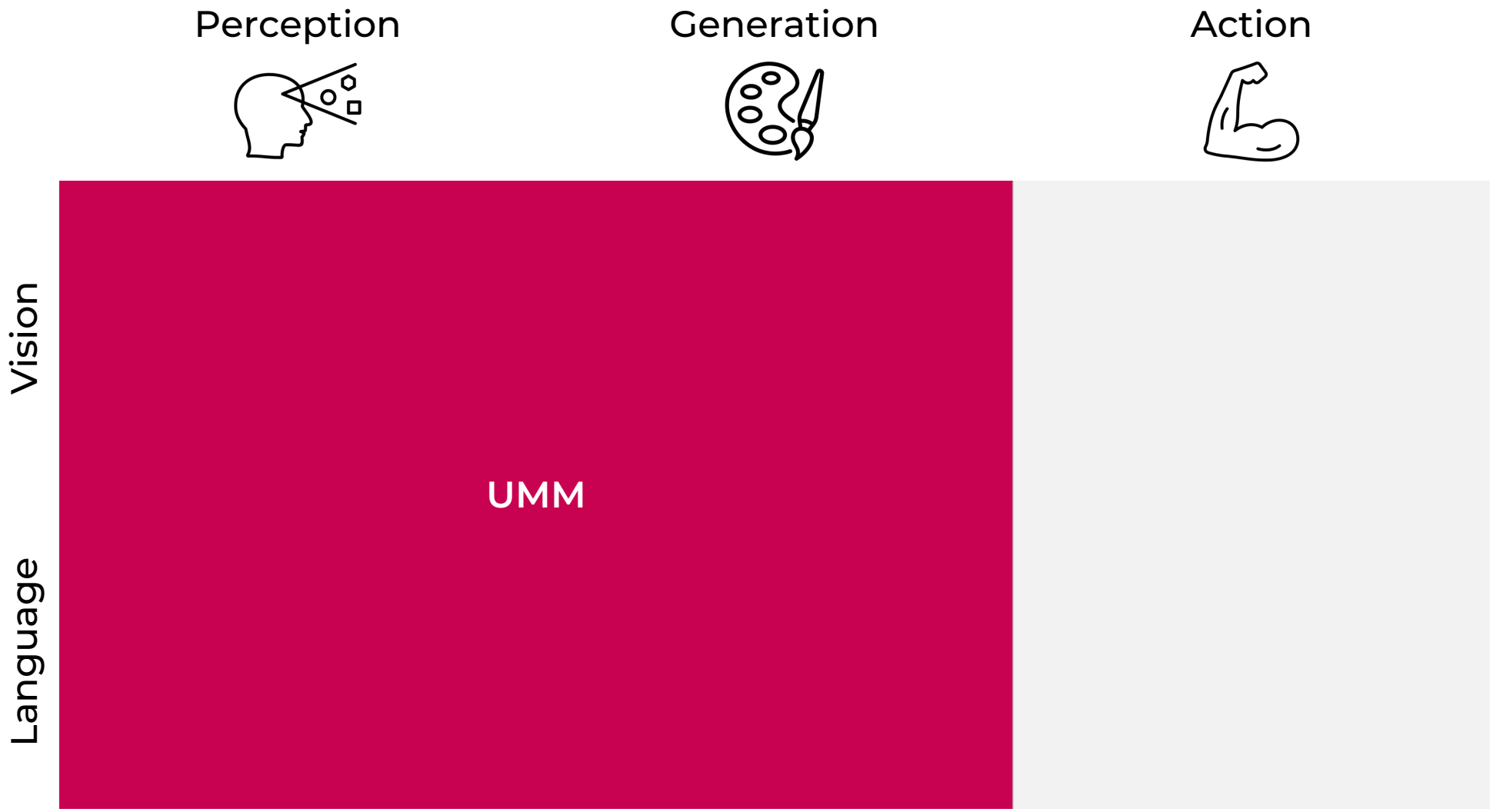
Visual Generation

	Perception 	Generation 	Action 
Vision	MLLM	Diffusion	
Language	MLLM	MLLM	

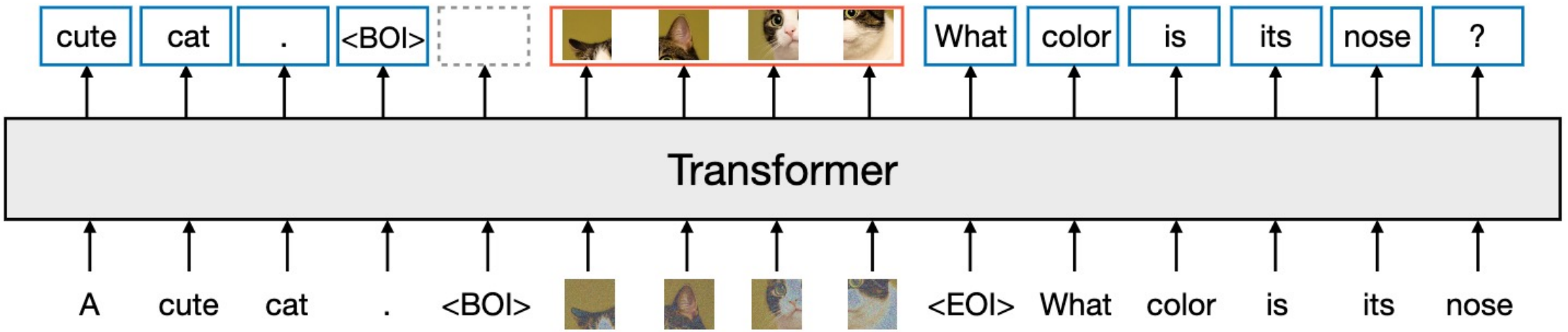
MLLM + Visual Generation?



Unified Multimodal Models

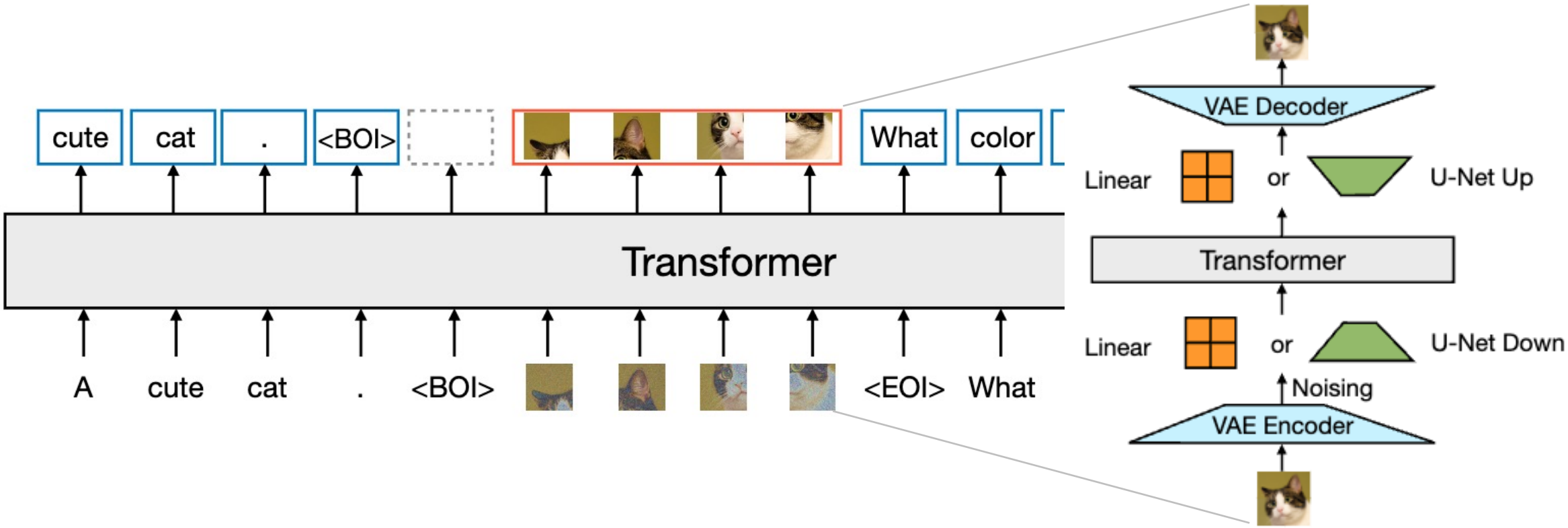


Unified Multimodal Models



Zhou et al. Transfusion: Predict the Next Token and Diffuse Images with One Multi-Modal Model. ICLR 2025.

Unified Multimodal Models

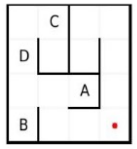


Zhou et al. Transfusion: Predict the Next Token and Diffuse Images with One Multi-Modal Model. ICLR 2025.

Unified Model: Visual Chain-of-Thought (CoT)

Input

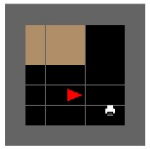
Maze



Action Sequence:

Go left. Go left. Go up. Go left. Go up.

MiniBehavior



Action Sequence:

Go right. Go right. Pick up. Go left. Go left. Go up. Drop.

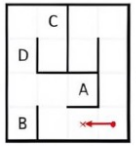
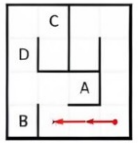
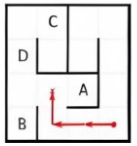
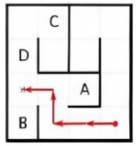
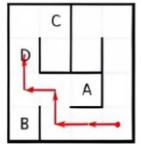
FrozenLake

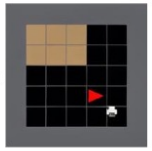
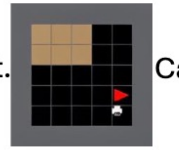
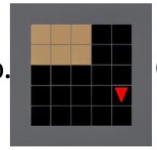
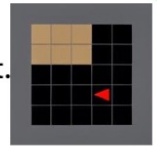


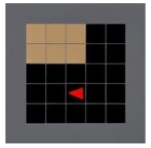
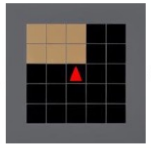
Action Sequence:

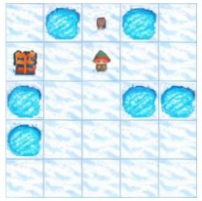
Go down. Go down. Go right. Go right. Go up. Go left. Go left.

Generated Response

Go left.  Go left.  Go up.  Go left.  Go up.  **The answer is D.**

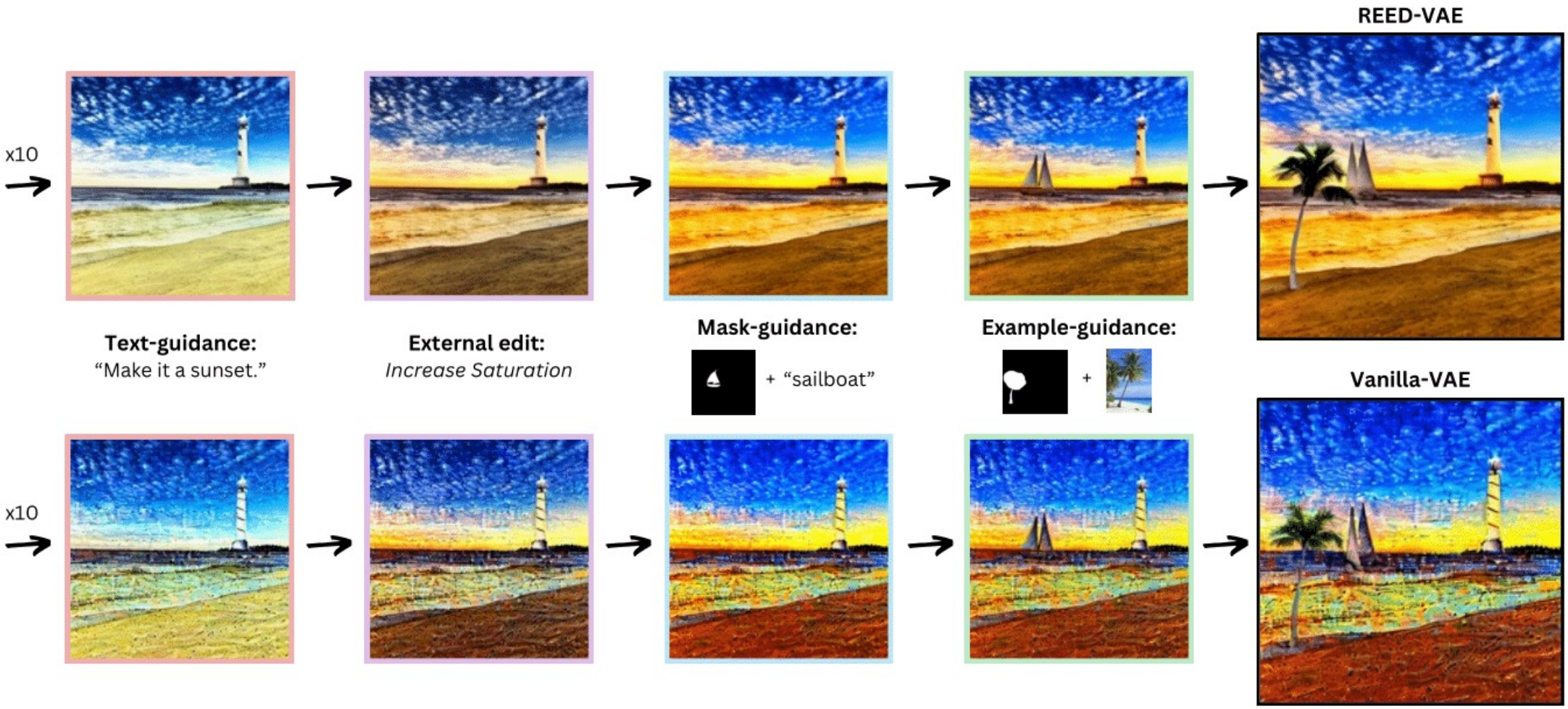
Go right.  Carrying: None. Go right.  Carrying: None. Pick up.  Carrying: printer. Go left. 

Carrying: printer. Go left.  Carrying: printer. Go up.  Carrying: printer. Drop.  Carrying: None. **Action Success.**

Go down.  Go down.  Go right.  **Action Failed: Fall into the Hole.**

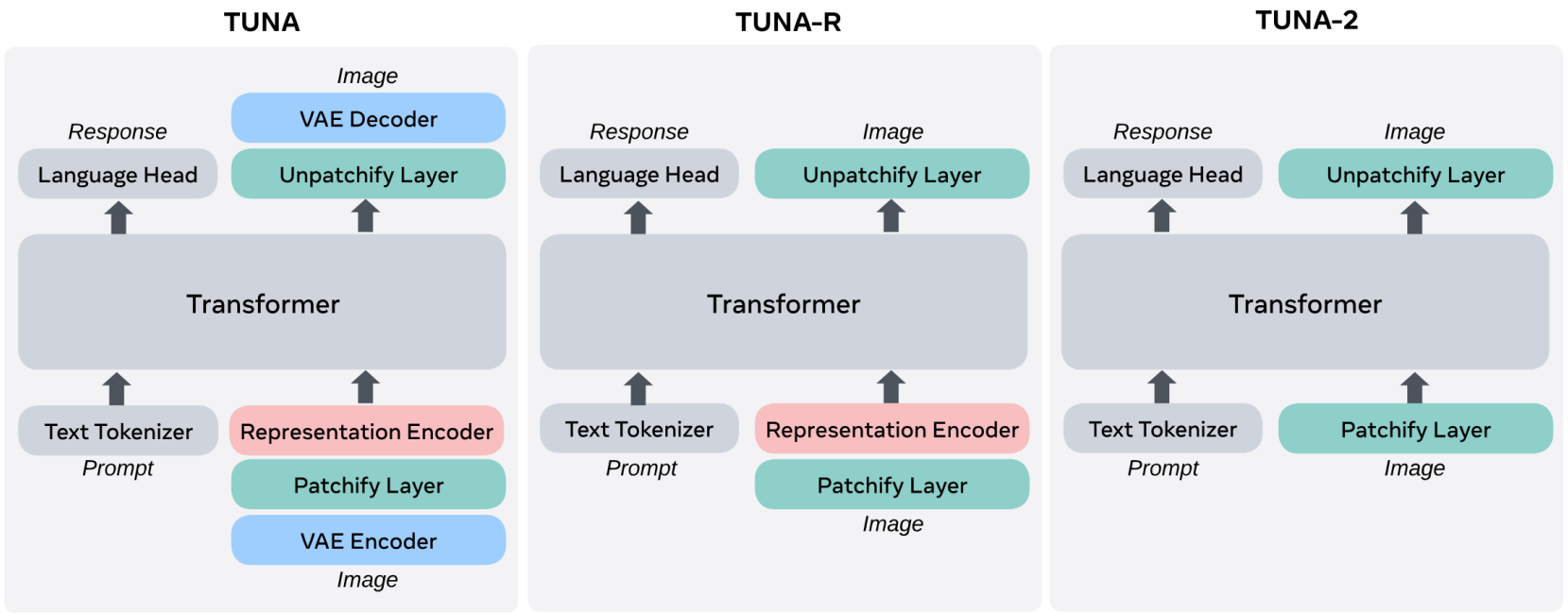
Li et al. Imagine while Reasoning in Space: Multimodal Visualization-of-Thought. arXiv 2025.

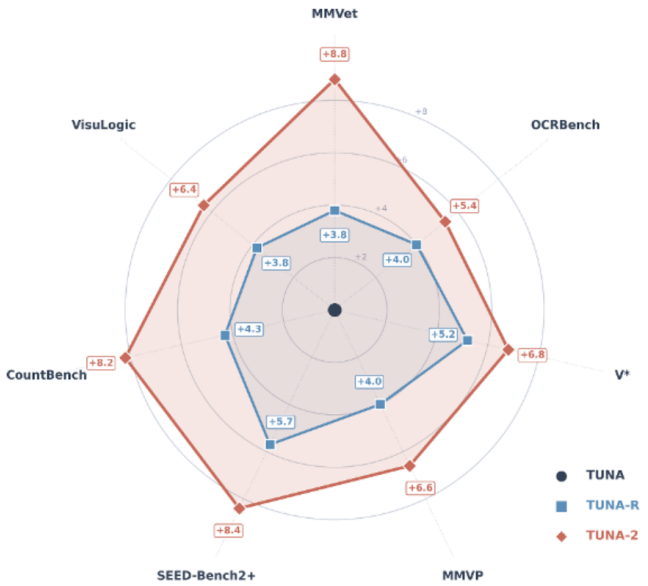
Unified Model: Iterative Image Editing



Amog et al. REED-VAE: RE-Encode Decode Training for Iterative Image Editing with Diffusion Models. arXiv 2025.

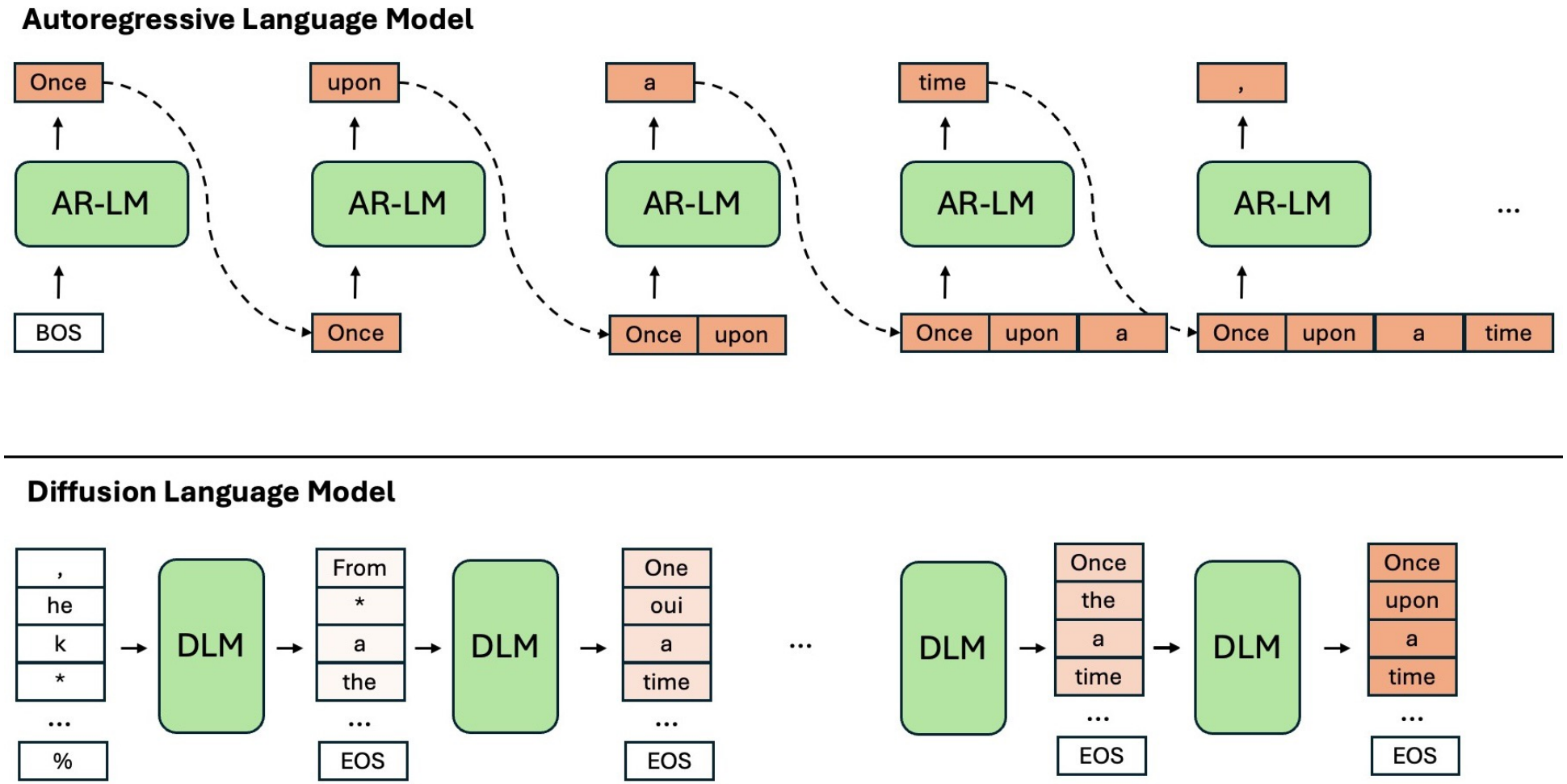
Q1. Do We Need Encoders?





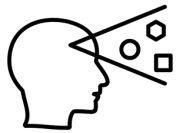
Liu et al. Tuna-2: Pixel Embeddings Beat Vision Encoders for Multimodal Understanding and Generation. arXiv 2026.

Q2. Autoregression vs. Diffusion



Hu et al. ELF: Embedded Language Flows. arXiv 2026.

Progress So Far – Part V

	Perception	Generation	Action
			
Vision	Visual Perception ⊂ UMM	Visual Generation ⊂ UMM	
Language	Lang. Understanding ⊂ UMM	Lang. Generation ⊂ UMM	