

Language Technologies Institute

Carnegie Mellon University

Multimodal Machine Learning

Lecture 12.1: Generation - Coherence

Louis-Philippe Morency

** Original course co-developed with Tadas Baltrusaitis. Major revision co-developed with Paul Liang. Co-lecturers included Yonatan Bisk, Daniel Fried and Carlos Busso.*

Administrative Stuff

Upcoming Course Assignments

Week 12 Reading Assignment (due Tuesday 11/17 at 8pm ET)

- Two papers (with a focus on multimodal generation)

Final project presentations (Tuesday 12/1 and Thursday 12/3)

- Read final report instructions first!
- Poster presentations
  - Poster will focus on a subset of the research ideas and results

Final project report (due Sunday 12/6 at 8pm ET)

- N-1 research ideas, exploring new modeling approaches
  - All ideas should be detailed in the report

Advanced Multimodal Machine Learning (11-877)

11-877 AMML

home schedule syllabus projects



Advanced Topics in MultiModal Machine Learning

11-877 • Spring 2024 • Carnegie Mellon University

Multimodal machine learning (MMML) is a vibrant multi-disciplinary research field which addresses some of the original goals of artificial intelligence by integrating and modeling multiple communicative modalities, including language, vision, and acoustic. This research field brings some unique challenges for multimodal researchers given the heterogeneity of the data and the contingency often found between modalities. This course is designed to be a graduate-level course covering recent research papers in multimodal machine learning, including technical challenges with representation, alignment, reasoning, generation, co-learning and quantification. The main goal of the course is to increase critical thinking skills, knowledge of recent technical achievements, and understanding of future research directions.

https://multicomp.cs.cmu.edu/adv-mmml-course/spring2024/

Attendance using Poll Everywhere



1. Scan QR Code
   or
   https://pe.app/mmml

2. Enter your AndrewID
   (as your “name tag”)

❌ *Do not try to sign-in*

3. Click ”Share location”



Language Technologies Institute

Carnegie Mellon University

Multimodal Machine Learning

Lecture 12.1: Generation – Coherence

Louis-Philippe Morency

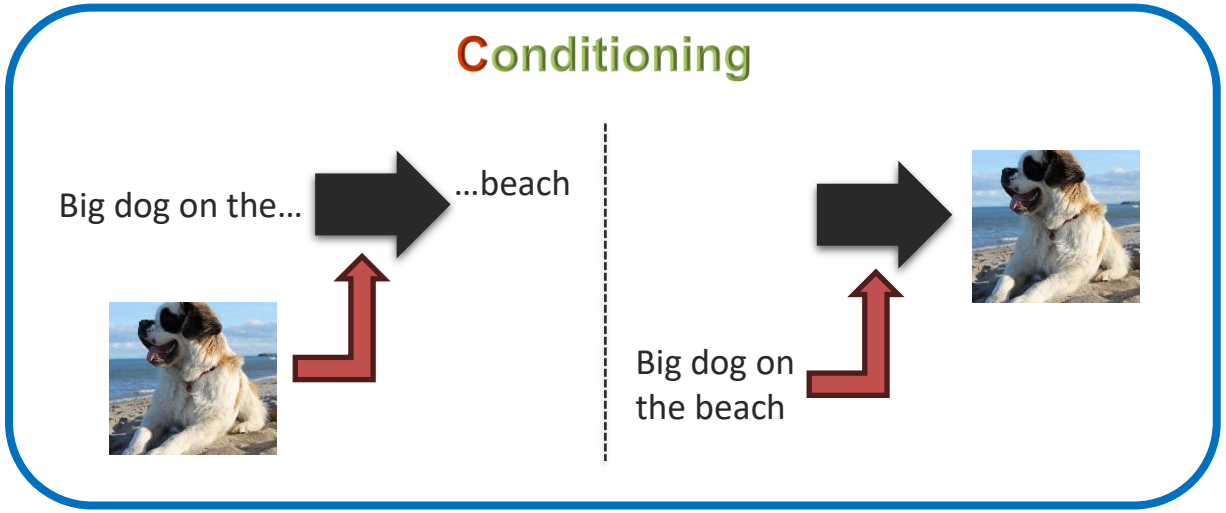
** Original course co-developed with Tadas Baltrusaitis. Major revision co-developed with Paul Liang. Co-lecturers included Yonatan Bisk, Daniel Fried and Carlos Busso.*

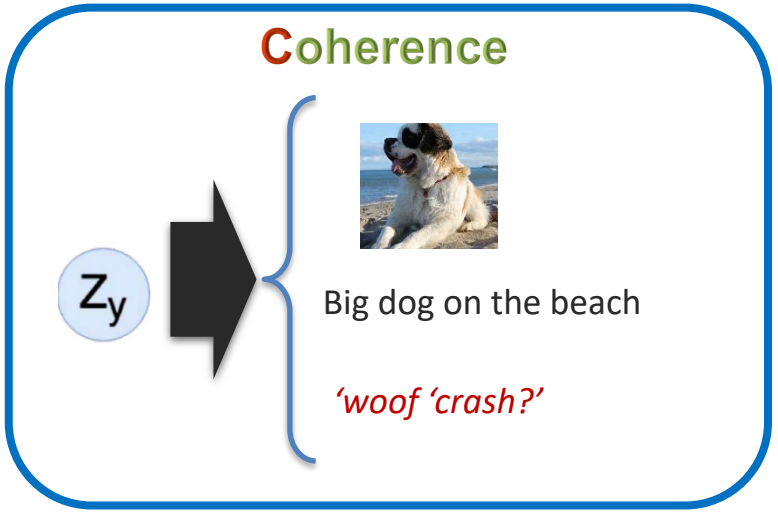
Objectives of today's class

- Coherence: generating several modalities from one model
- Discrete image tokens: VQ-VAE and VQ-GAN
- One vocabulary for text and images (Chameleon)
  - Modality competition and training stability
- Separate encoders for understanding and generation (Janus)
- One transformer, two losses (Transfusion)
- Streaming generation of text and speech (Qwen-Omni)
  - Temporal alignment of audio and video (TM-RoPE)

Generation Challenge (v2)

Definition: Learning a generative process to produce raw modalities that reflects cross-modal interactions, structure, and coherence.

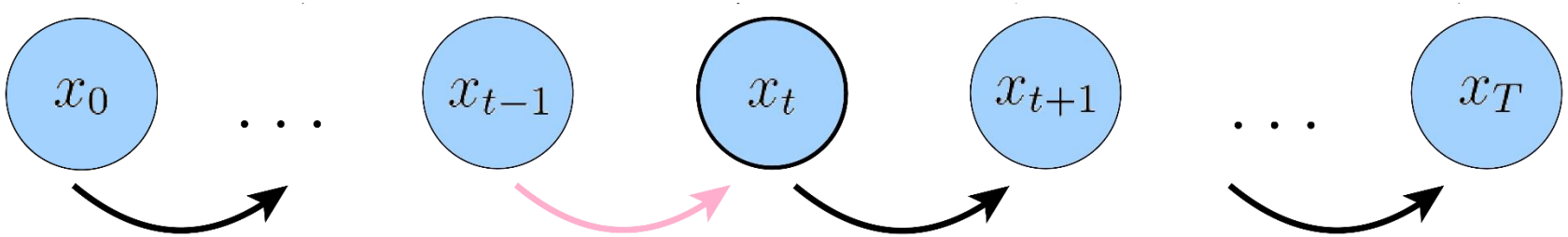




Conditioning for Diffusion-based Generation

Diffusion: Corrupting Images to Learn how to Generate Them

Corruption: a blend of the clean image and pure noise



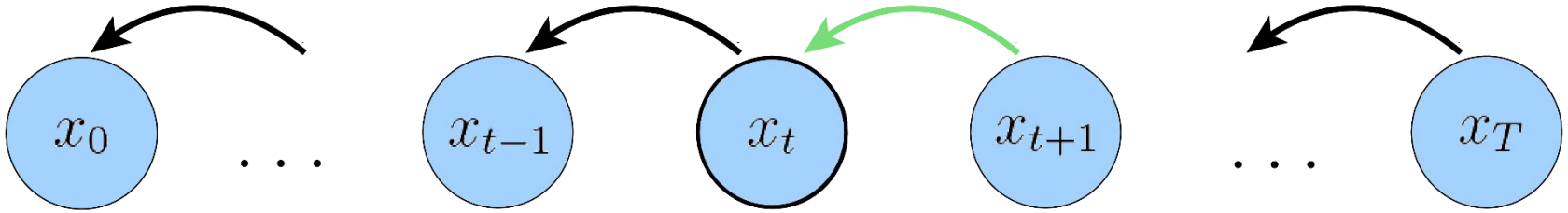


$$x_t = \underbrace{\sqrt{\bar{\alpha}_t}}_{\text{signal}} x_0 + \underbrace{\sqrt{1-\bar{\alpha}_t}}_{\text{noise}} \varepsilon \quad \text{where } \varepsilon \sim \mathcal{N}(0, I)$$

[Ho, Jain & Abbeel, Denoising Diffusion Probabilistic Models. NeurIPS 2020]

Reverse process: Image Generation

Question: Which image x_0 produced this noisy image x_t?



Key Idea: If each corruption step is small, the reverse step is also approximately Gaussian



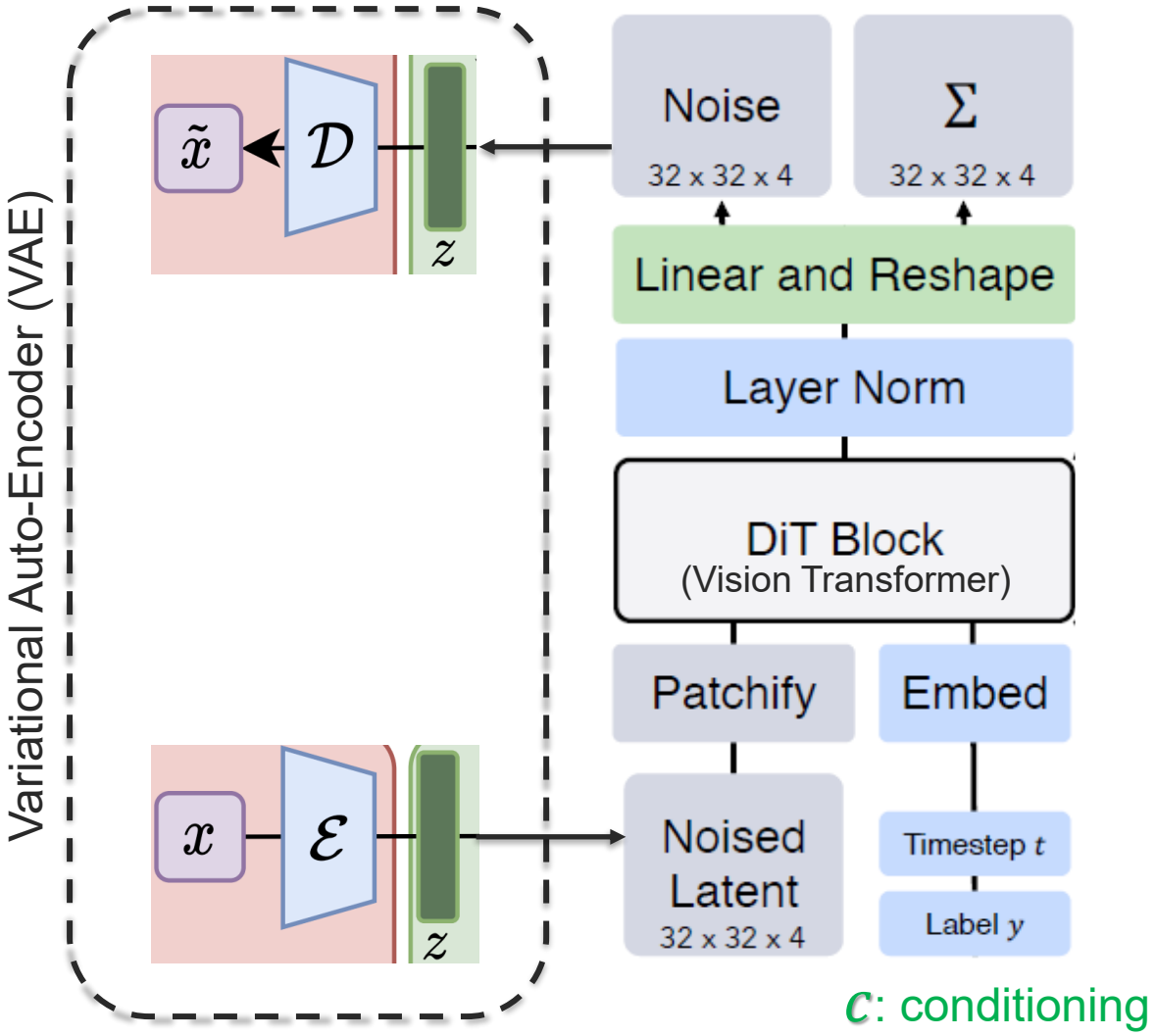
Then, let's learn a network that can predict the noise

$$\hat{\varepsilon} = \varepsilon_\theta(x_t, t)$$

$$\hat{x}_0 = \frac{x_t - \sqrt{1-\bar{\alpha}_t}\hat{\varepsilon}}{\sqrt{\bar{\alpha}_t}}$$

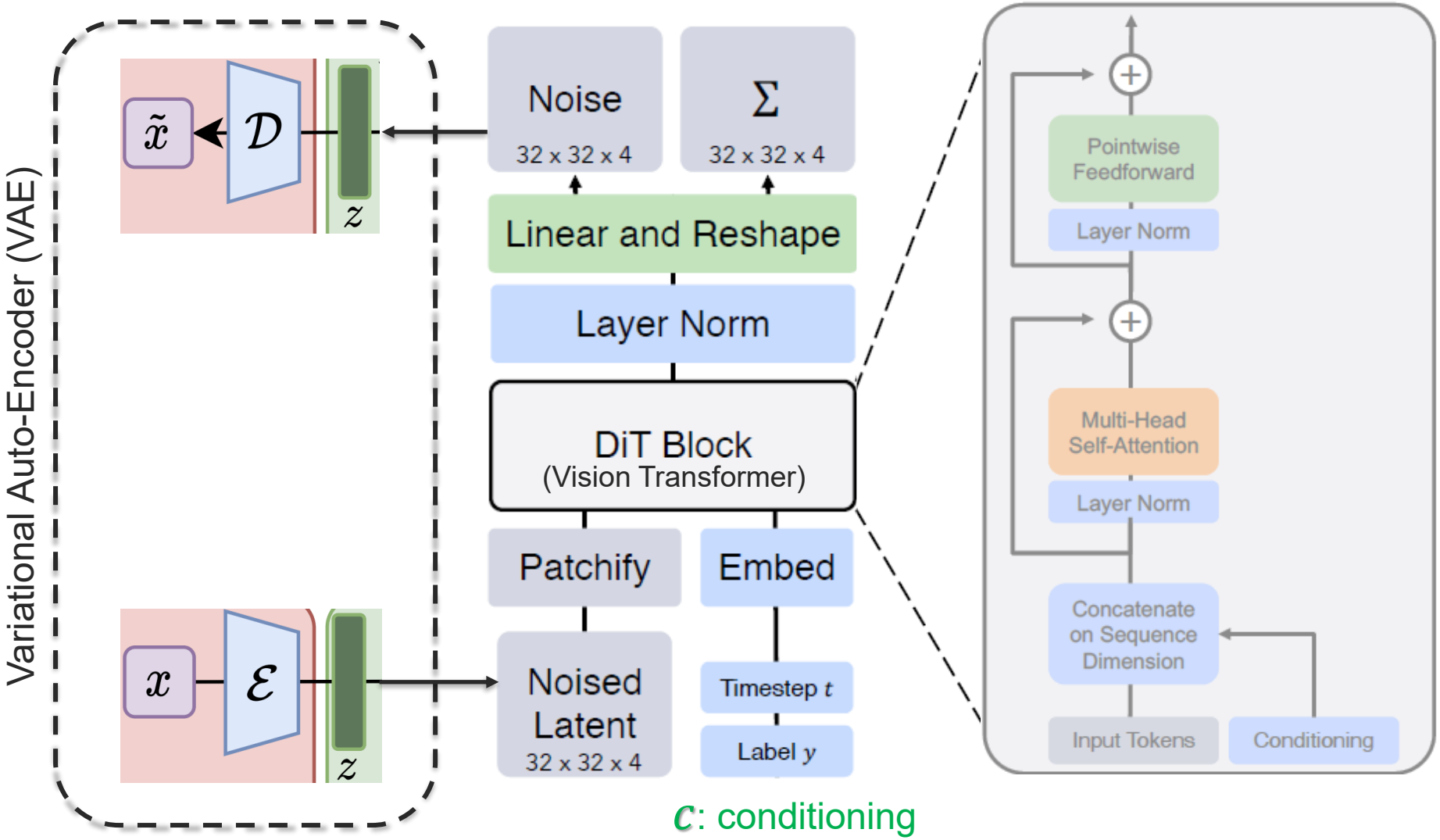
[Ho, Jain & Abbeel, Denoising Diffusion Probabilistic Models. NeurIPS 2020]

Diffusion Transformer



[Peebles & Xie, Scalable Diffusion Models with Transformers. ICCV 2023]

Conditioning Strategy A: In-Context Conditioning



$\varepsilon_\theta(z_t, t, c)$

Conditioning simply added as extra tokens

[Peebles & Xie, Scalable Diffusion Models with Transformers. ICCV 2023]

Classifier-Free Guidance

At training time:

$$c \leftarrow \emptyset \quad \text{with probability} \approx 0.1$$

At inference time:

Two forward passes: $\varepsilon_\theta(z_t, t, c)$ and $\varepsilon_\theta(z_t, t, \emptyset)$

and then combined: $\tilde{\varepsilon} = \varepsilon_\theta(z_t, t, \emptyset) + w\big(\varepsilon_\theta(z_t, t, c) - \varepsilon_\theta(z_t, t, \emptyset)\big)$

guidance scale

[Ho & Salimans, Classifier-Free Diffusion Guidance. arXiv 2022]

Where Does the Condition Enter?

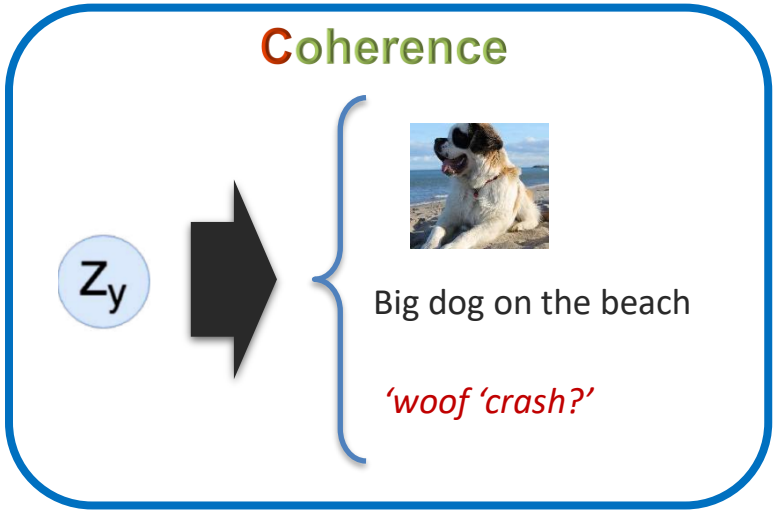
Where	How	Where you saw it
The input	extra tokens in the sequence	Strategy A — MM-DiT
The hidden layers	cross-attention	Strategy B — Latent Diffusion
	scale & shift	Strategy C (AdaLN) — DiT, MM-DiT
	a side network adds features	ControlNet
The output	conditional - unconditional, scaled by w	Classifier-free guidance

Note: Also possible: edit z_t between steps, with no training at all (Attend-and-Excite, 2023)

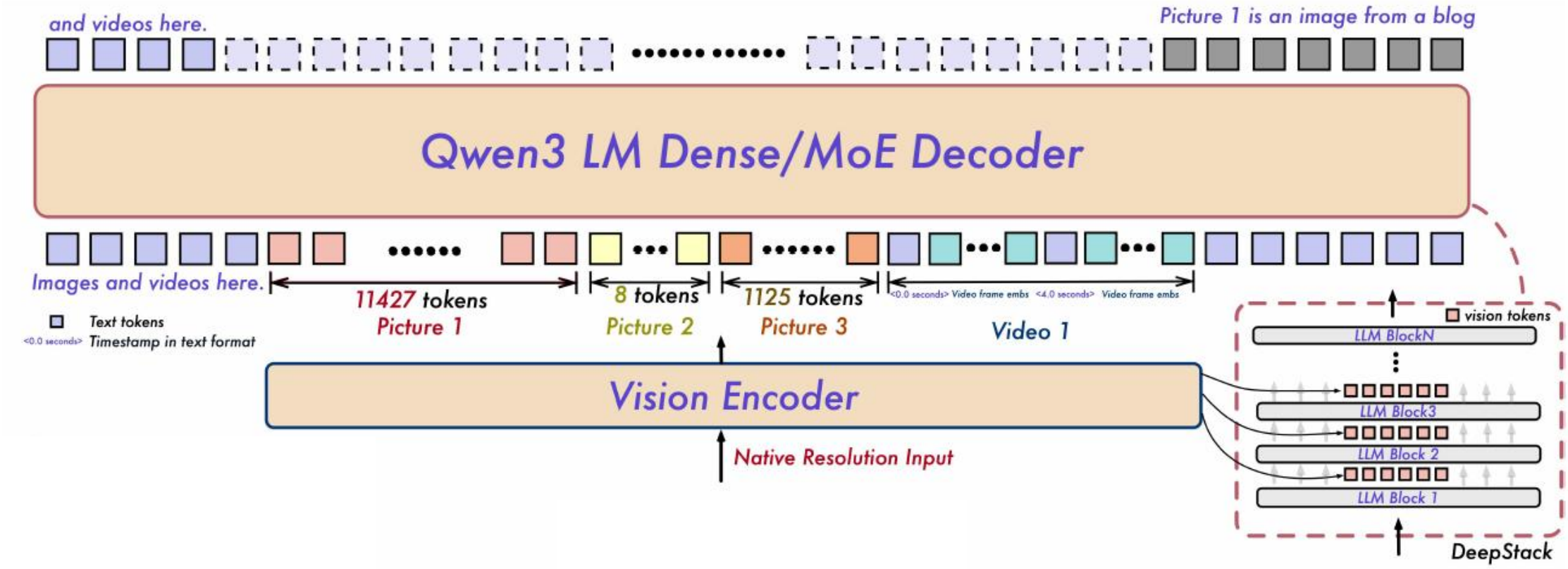
Coherence

Coherence Sub-Challenge

Definition: Learning a generative process to produce raw modalities that reflects cross-modal interactions, structure, and coherence.

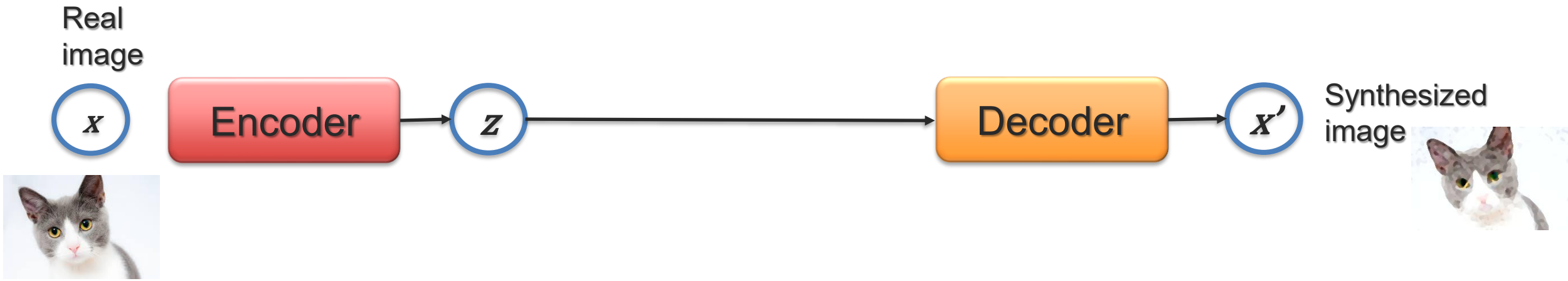


Reminder: Visual + Text Inputs (VLMs and MLLMs)



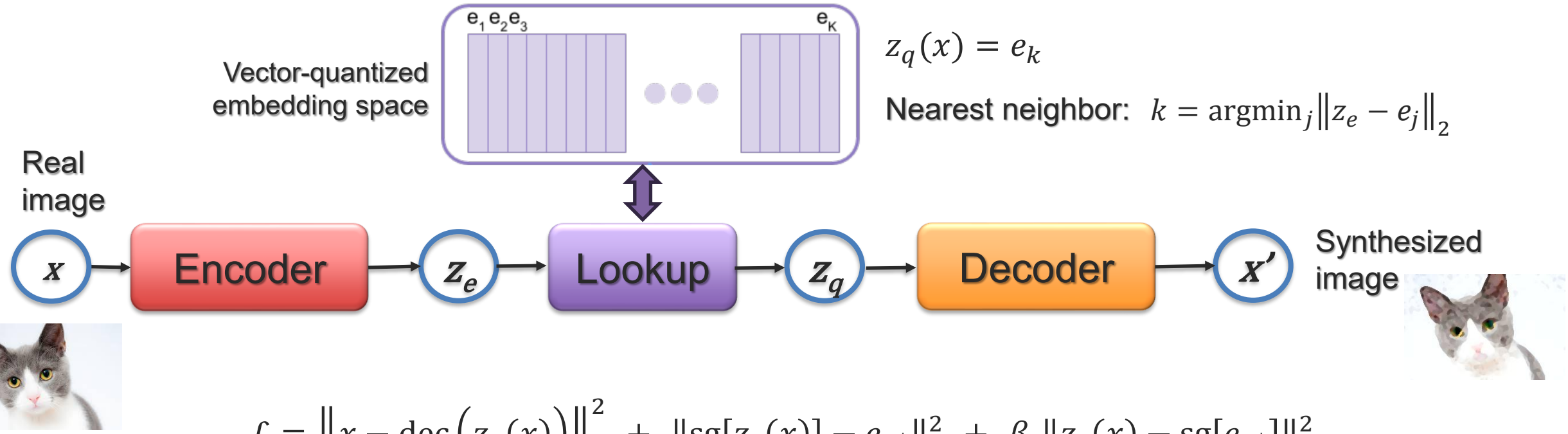
But how to also generate visual outputs with autoregressive model?

Auto-encoder



[Kingma & Welling, Auto-Encoding Variational Bayes. ICLR 2014]

Vector-Quantized VAE (VQ-VAE)

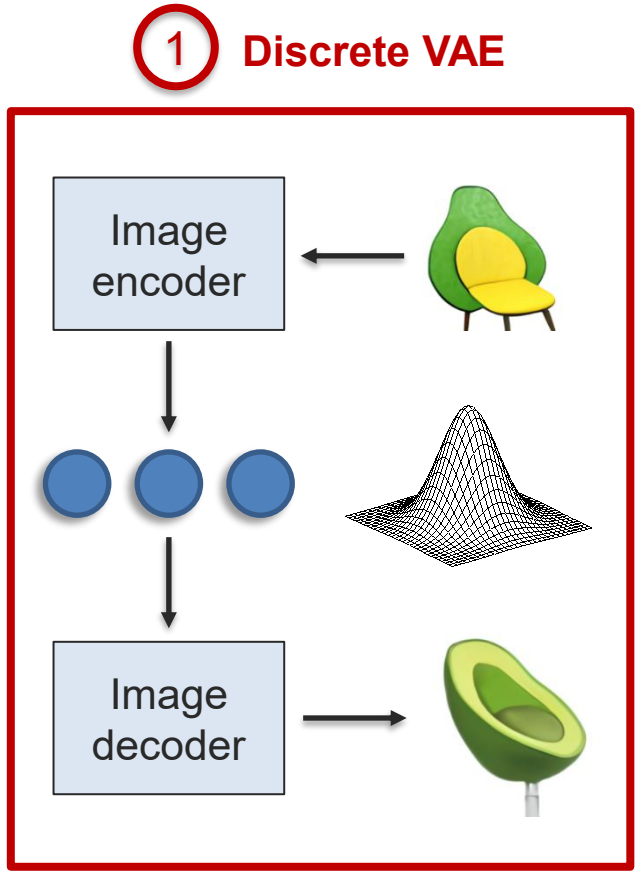


$$\mathcal{L} = \underbrace{\left\| x - \text{dec}\left(z_q(x)\right)\right\|_2^2}_{\textbf{reconstruction}} + \underbrace{\|\text{sg}[z_e(x)] - e_{k^*}\|_2^2}_{\textbf{codebook}} + \beta \underbrace{\|z_e(x) - \text{sg}[e_{k^*}]\|_2^2}_{\textbf{commitment}}$$

Encoder & Decoder | Lookup | Encoder

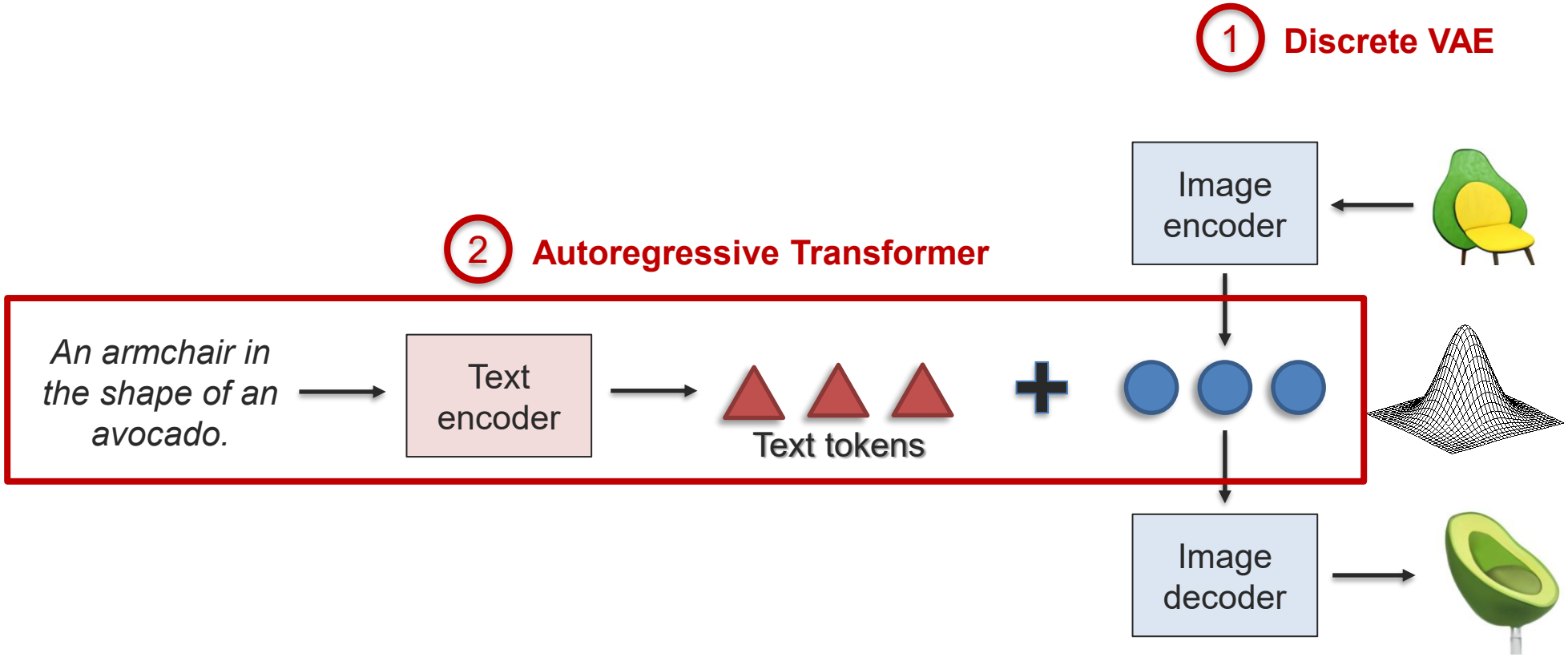
[van den Oord et al., Neural Discrete Representation Learning, NIPS 2017]

DALL·E (v1): Text-to-image translation at scale



[Ramesh et al., Zero-Shot Text-to-Image Generation. ICML 2021]

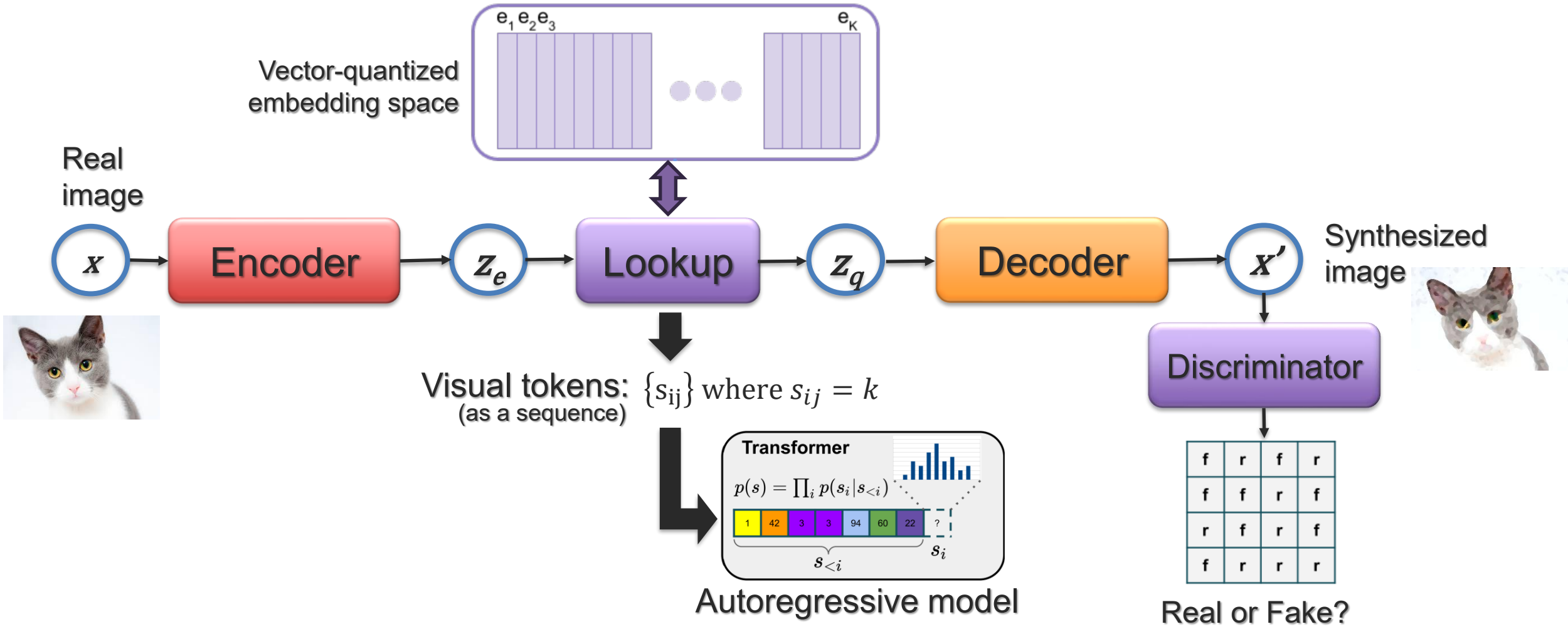
DALL·E (v1): Text-to-image translation at scale



Text tokens, then image tokens — always in that order.

[Ramesh et al., Zero-Shot Text-to-Image Generation. ICML 2021]

VQ Generative-Adversarial Network (VQ-GAN)



[Esser et al., Taming Transformers for High-Resolution Image Synthesis, 2021]

Mixed-Modal: Going Beyond Text-to-Image Translation



PROMPT: What can I bake with this? <img>

Prompt: I'm bored. Could you show me some cool, quirky-looking birds? Also tell me about them in 2–3 sentences.



Response: Absolutely! Here are three birds with some unique features that you might find interesting:

Keel-Billed Toucan The Keel-Billed Toucan is a brightly colored bird native to South American rainforests and woodlands. Its most distinctive feature is its colorful beak, which can grow up to seven inches long and has a distinctive, curved shape. The beak is used for plucking fruit from trees and for social displays. Despite its large size, the Keel-Billed Toucan is a skilled flyer and can often be seen flitting through the treetops. Here is an image of this vibrant bird: <img>

Puffin Puffins are small, aquatic birds that are found in the northern hemisphere. They have distinctive black and white feathers, as well as a brightly colored beak that is used for attracting mates and defending their territory. Puffins are known for their distinctive "puffin" call, which sounds like a loud, guttural chuckle. They feed on fish, which they catch by diving into the water from the air. Puffins are colonial birds, and they breed in large colonies on rocky cliffs and islands. Here is an image of this adorable bird: <img>

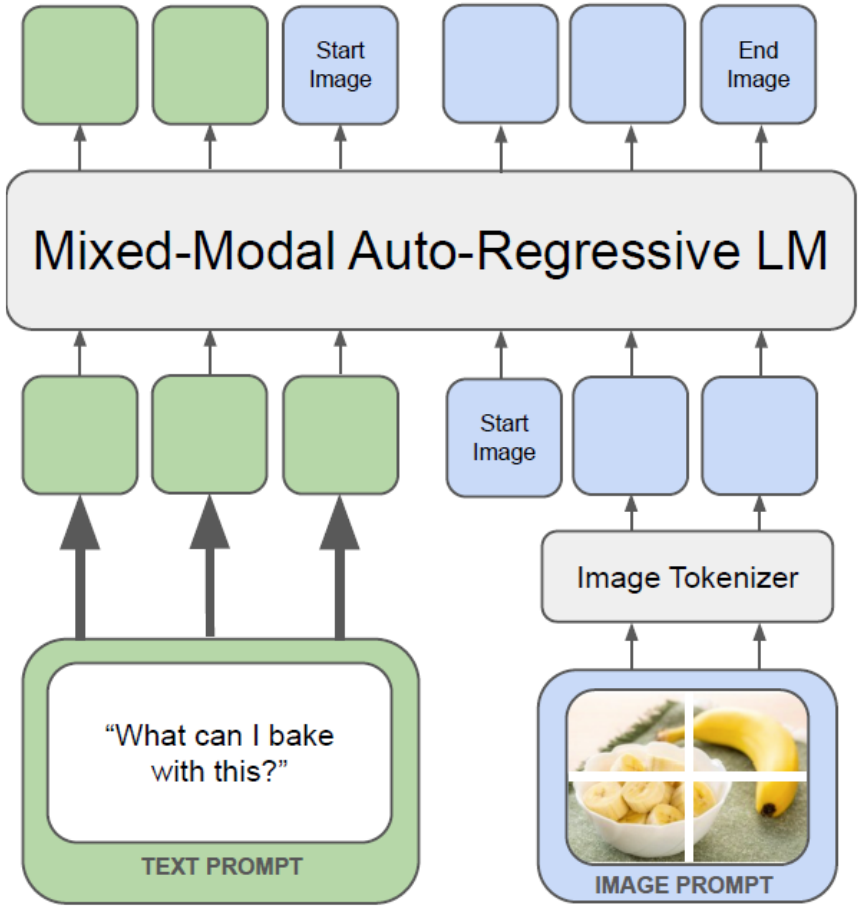




Golden Pheasant The Golden Pheasant is a brightly colored bird native to the forests of China. It is a member of the pheasant family and is known for its striking plumage, which features a golden-yellow body, red face, and green tail feathers. The Golden Pheasant is a ground-dwelling bird, and it feeds on a diet of insects, seeds, and fruits. It is a popular game bird, and is often hunted for its meat and feathers. In captivity, the Golden Pheasant is sometimes kept as an ornamental bird, due to its striking appearance. Here is an image of this beautiful bird: <img>

I hope these images and descriptions have helped to brighten your day! If you have any other questions or requests, please don't hesitate to ask.

Chameleon: Mixed-Modal Auto-Regressive LM



The vocabulary

- 8,192 image codes, 57,344 text tokens
- One vocabulary, one softmax, one loss

Mixed-modal documents

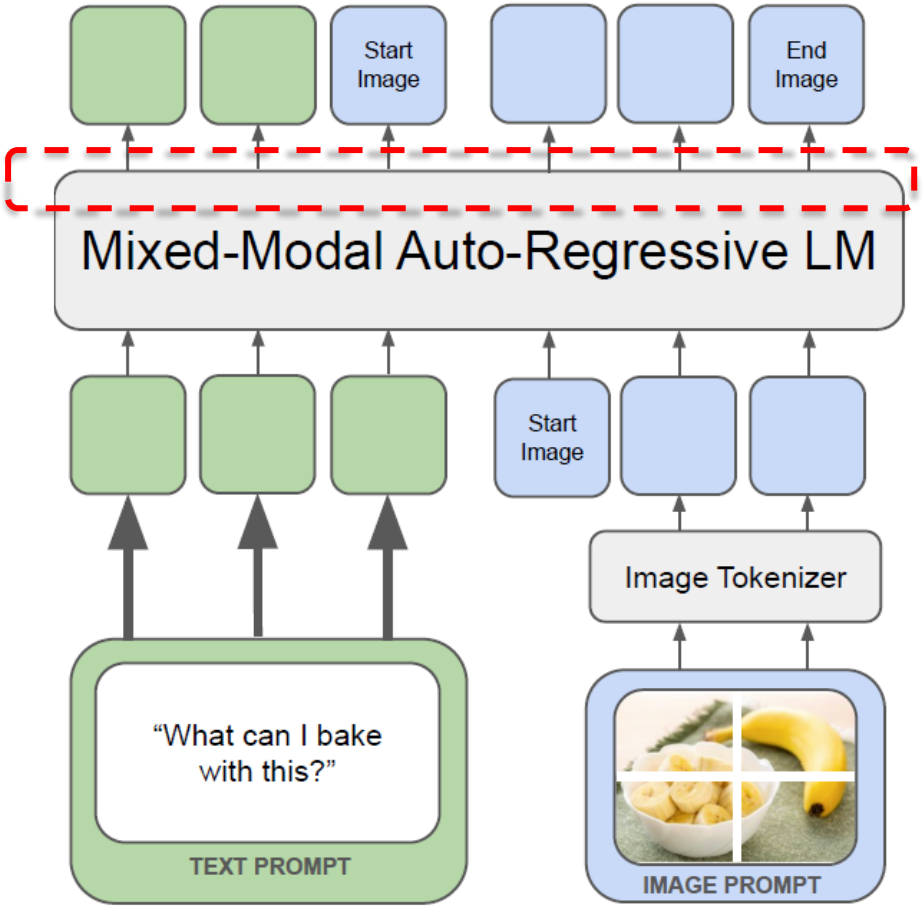
- Any order, any number of images
- Reads and writes both modalities

Training method

- 1.4B pairs → 1.5T image tokens
  - 400B interleaved tokens
- Early fusion from inception
  (no encoder, no adapter)

[Chameleon: Mixed-ModalEarly-FusionFoundation Models, 2024]

The Softmax Has a Blind Spot – The "Logit Drift" Problem



$$\text{softmax}([\boldsymbol{z}_{text}, \boldsymbol{z}_{visual}])$$

… but softmax is translation invariant:

$$\text{softmax}(\boldsymbol{z}) = \text{softmax}(\boldsymbol{z} + c)$$

Only logit differences matter!

During training, no gradient on the absolute level — nothing pulls it back

[Chameleon: Mixed-ModalEarly-FusionFoundation Models, 2024]

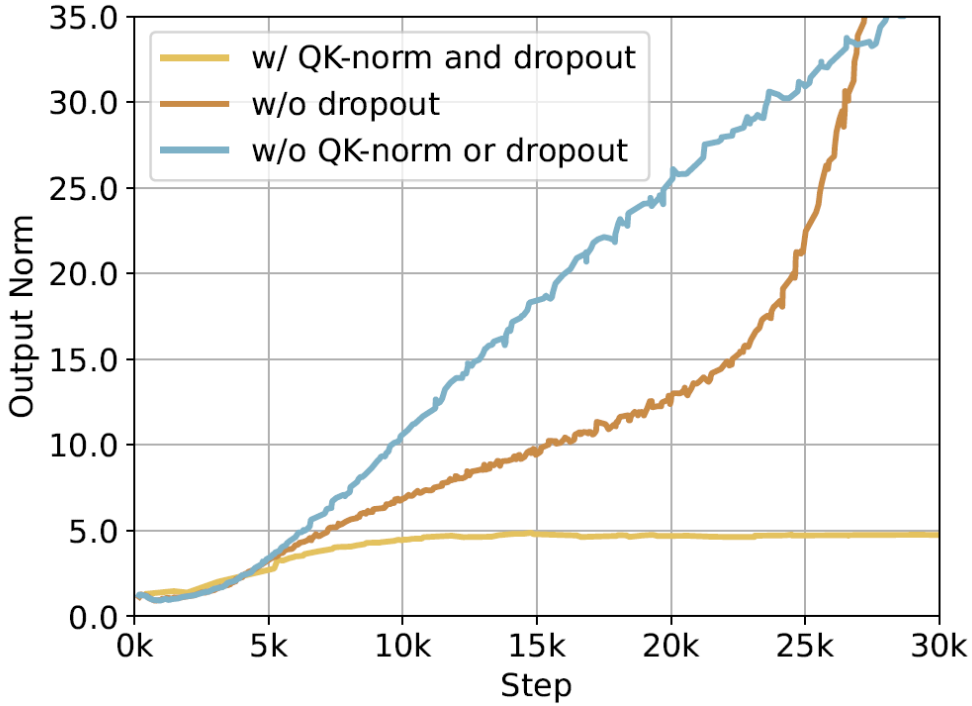
Modality Competition

"Because we share all weights of the model across modalities, each modality will try to 'compete' with the other by increasing its norms slightly"

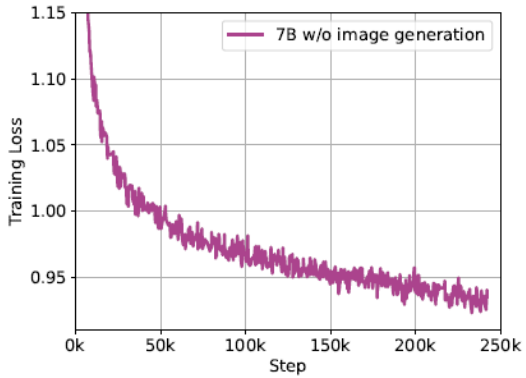
Without normalization, the norms grow…

…and high norms can saturate the attentions

Also, norm growth seems to predict future divergence.



Ablations without image generation did not diverge!



[Chameleon: Mixed-ModalEarly-FusionFoundation Models, 2024]

Stabilizing Multimodal Training

Solution 1: Normalize Q and K before the attention softmax

Standard attention: $Q = XW_Q, \quad K = XW_K$

QK-Norm: $Q = \mathrm{LN}(XW_Q), \quad K = \mathrm{LN}(XW_K)$

LN: LayerNorm

$$A = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V$$

Solution 2: Regularize the sum of the output logits

$$\mathbf{z-loss}: \quad \mathcal{L} \mathrel{+}= 10^{-5}\log^2 Z\,, \quad \text{where} \quad Z = \sum_j e^{z_j}$$

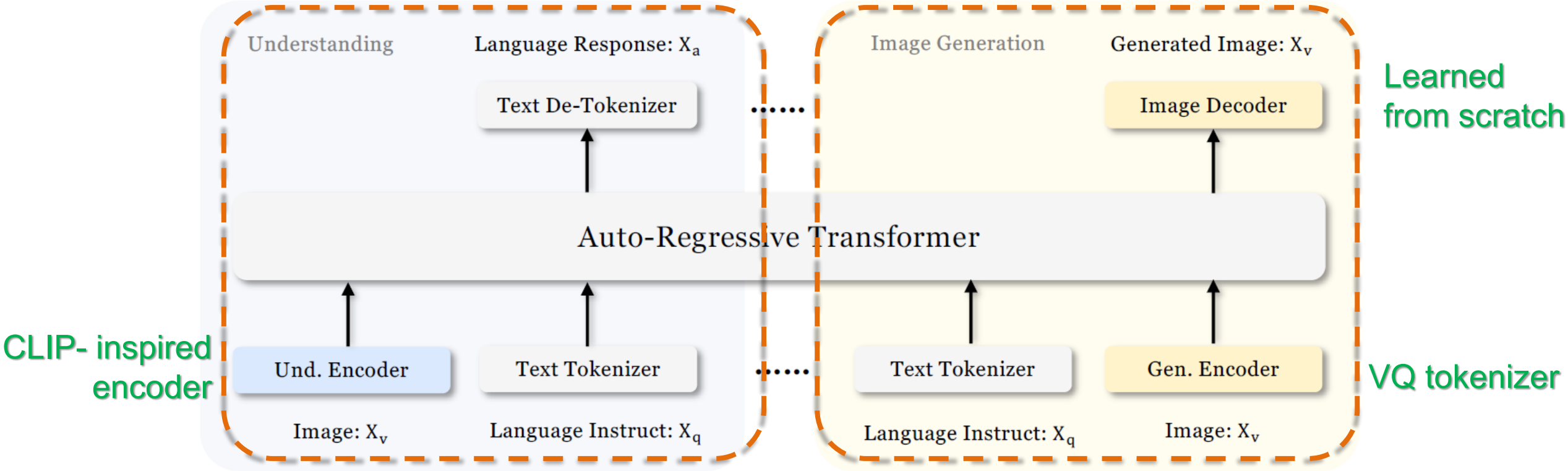
- Solution 1 is an architectural constraint: the norms *cannot* grow
- Solution 2 is a penalty: it discourages growth without preventing it

Good guardrails, but neither removes the cause…

[Chameleon: Mixed-ModalEarly-FusionFoundation Models, 2024]

Alternative Solution: Two Pipelines, One Shared Transformer

One encoder can't serve both: Understanding wants semantics. Generation wants detail.



One disadvantage: Not "mixed-modal" anymore. No images mid-sentence.

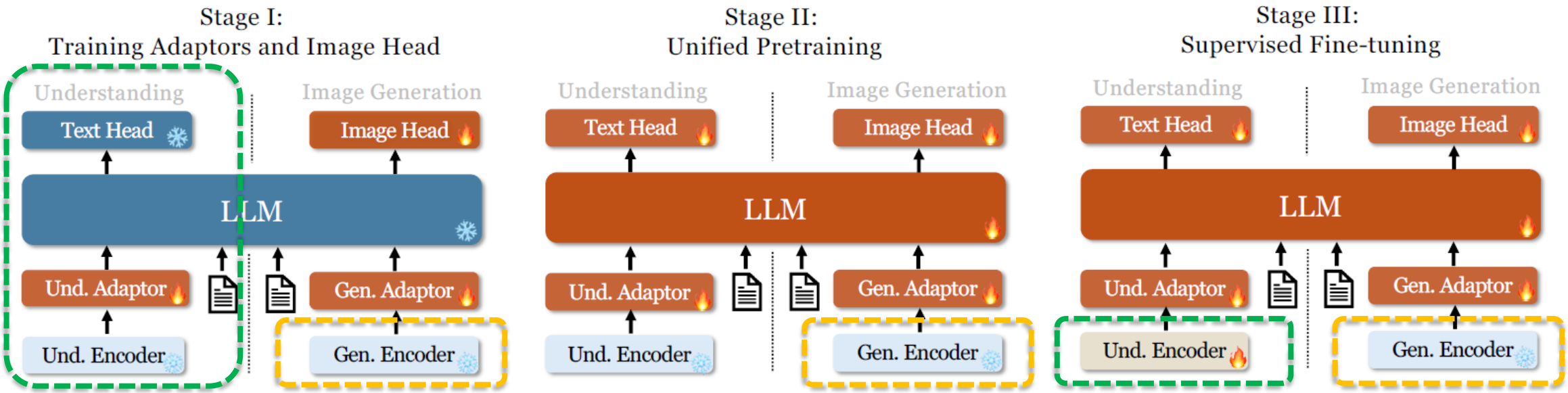
[Chen et al., Janus-Pro: Unified Multimodal Understanding and Generation with Data and Model Scaling, 2025]

Alternative Solution: Two Pipelines, One Shared Transformer

Align the connectors first
(similar to LLaVA)

LLM learns to generate

Instruction tuning
(similar to LLaVA)



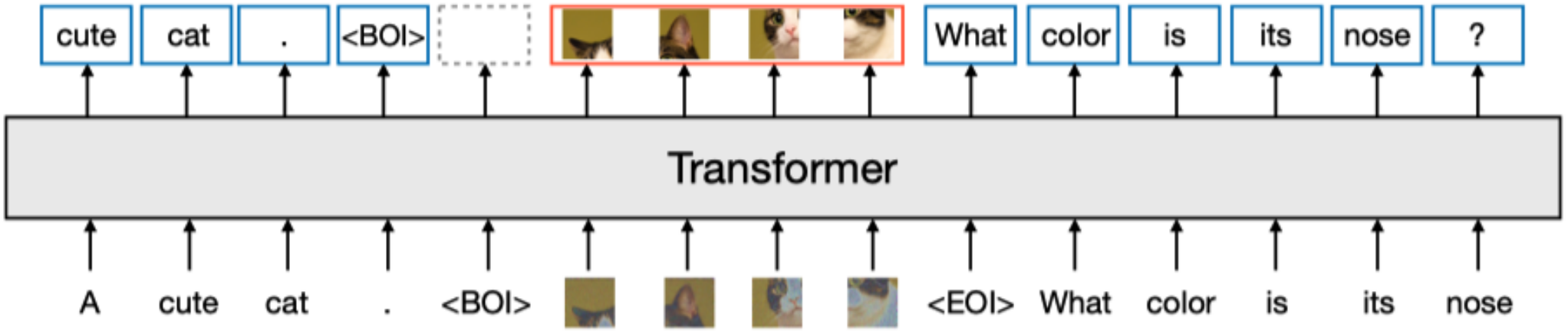
What VQ tokenizer discards, the model cannot recover

VQ tokenizer is never updated!
(a codebook is harder to update)

[Chen et al., Janus-Pro: Unified Multimodal Understanding and Generation with Data and Model Scaling, 2025]

Transfusion: One Transformer for Diffusion and Next-Token-Prediction

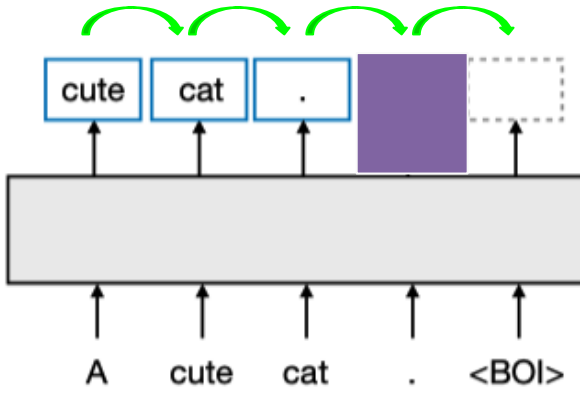
Next-token prediction loss **Diffusion loss** **Next-token prediction loss**



discrete text tokens, and **continuous image vectors**, In any order

Transfusion: One Transformer for Diffusion and Next-Token-Prediction

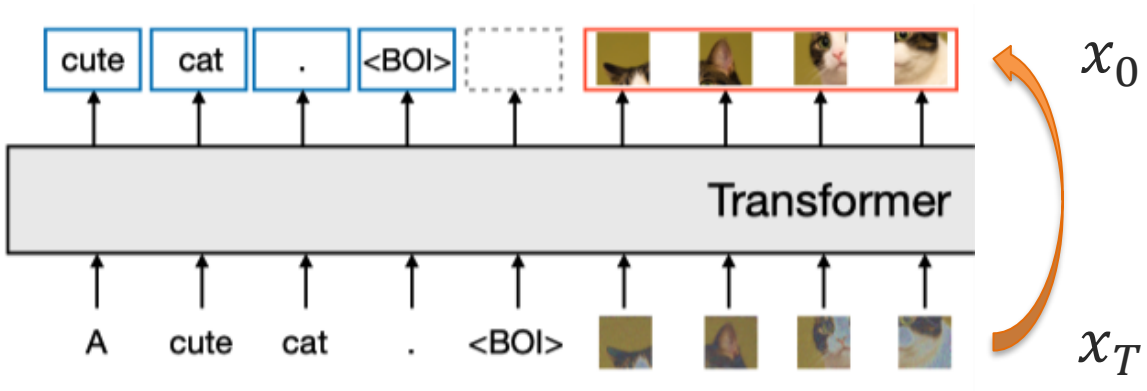
Next-token predictions



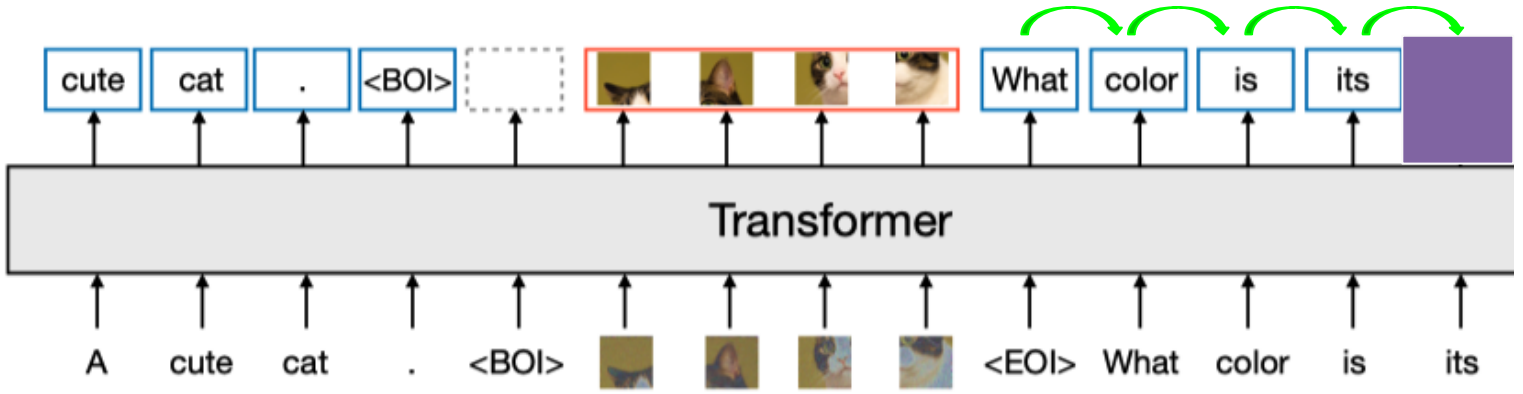
[Chen et al., Janus-Pro: Unified Multimodal Understanding and Generation with Data and Model Scaling, 2025]

Transfusion inference

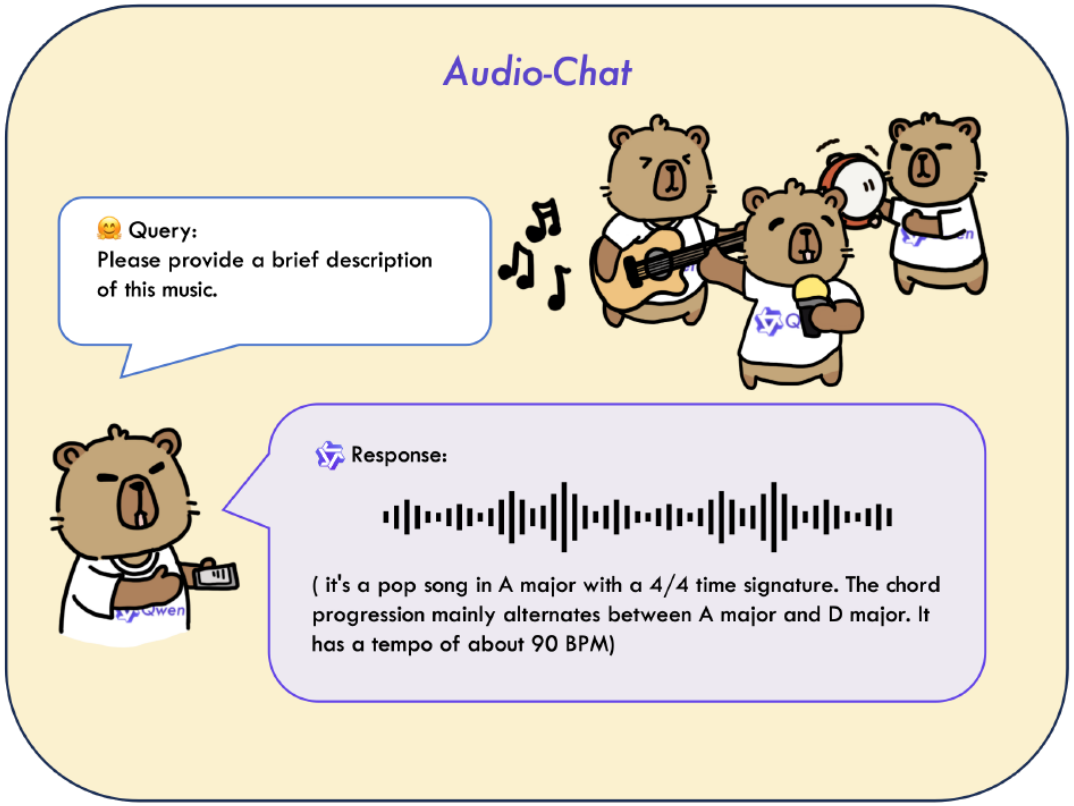
Inversed Diffusion inference

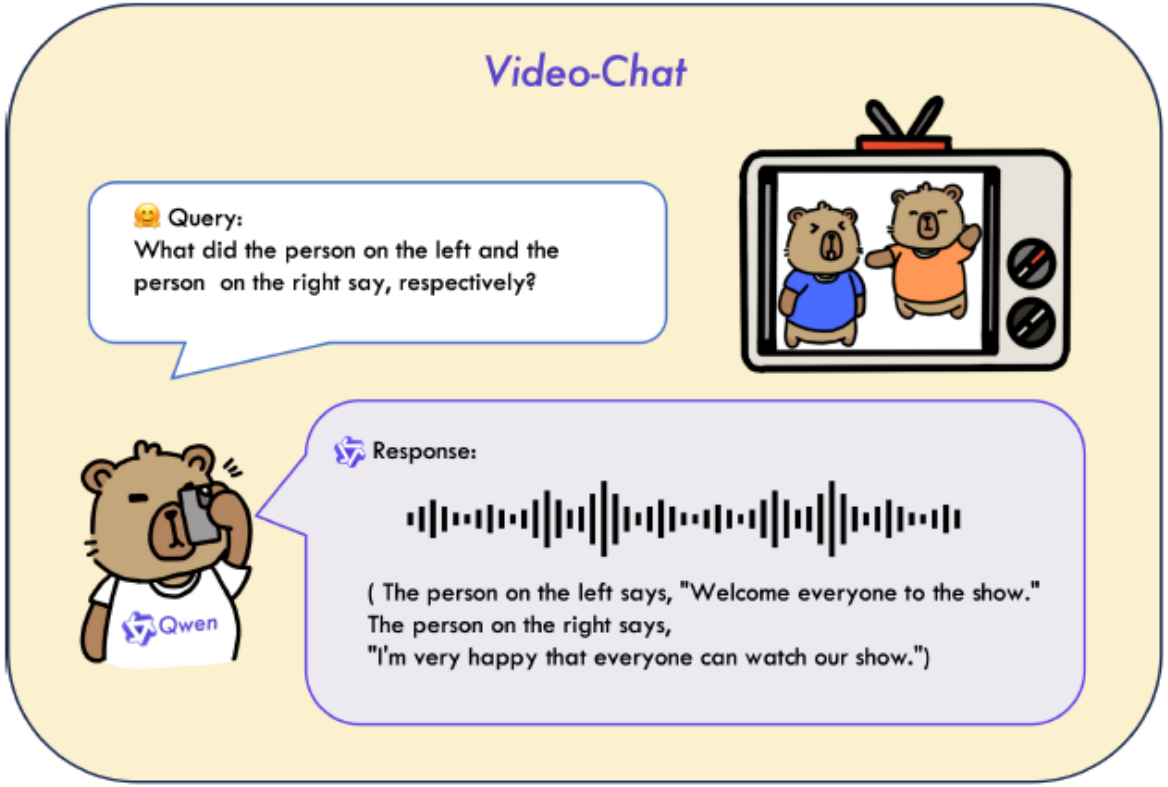


Transfusion inference



Adding Temporal Dimension (e.g., Videos and Audio Modalities)



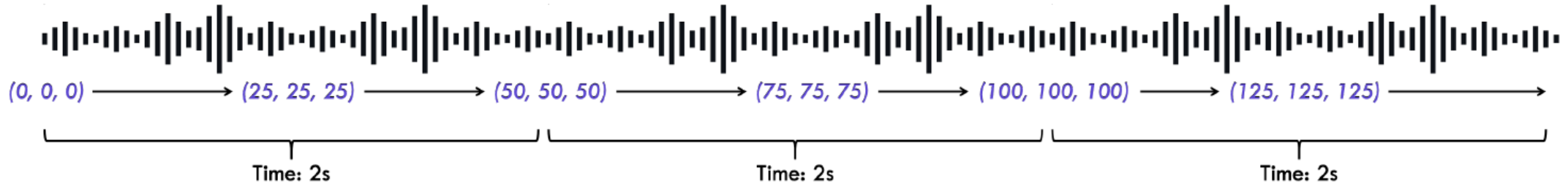


[Qwen2.5-Omni Technical Report, 2025]

Time-Aligned Multi-Dimensional RoPE (TM-RoPE)

Same three dimensions as M-RoPE: Temporal (t), Height (h), Width (w)

Token type	t	h	w
Text	sequence index	same	same
Image	constant (one frame)	patch row	patch column
Video	time / 40ms	patch row	patch column
Audio	time / 40ms	same	same



[Qwen2.5-Omni Technical Report, 2025]

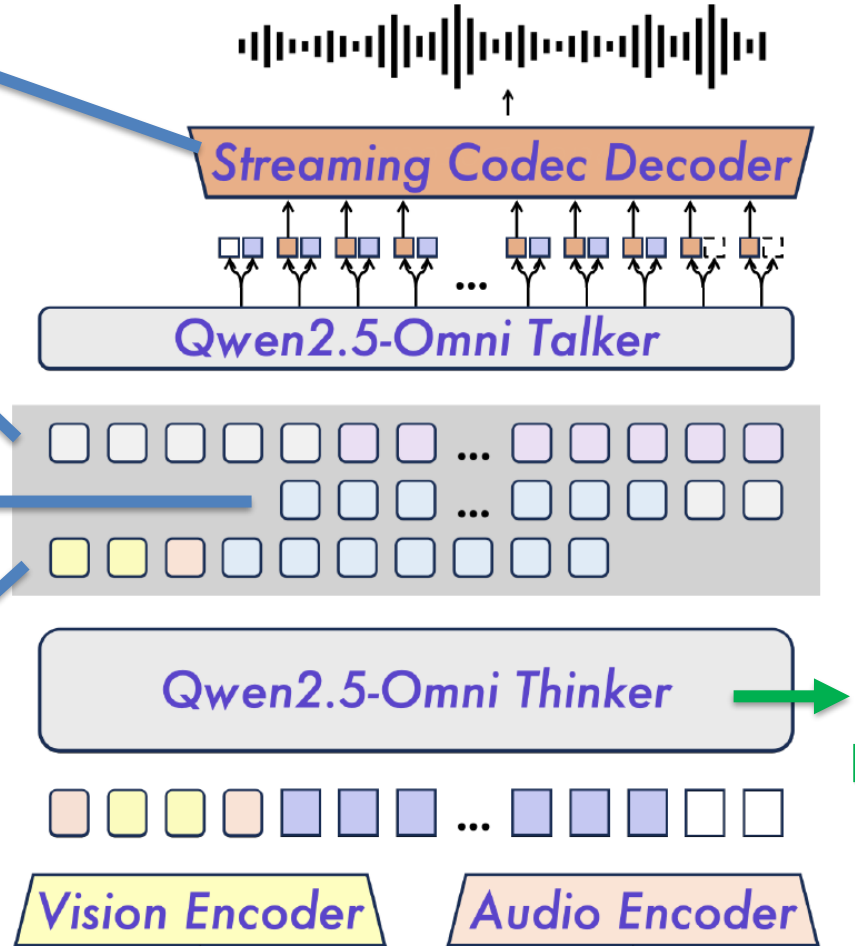
Qwen 2.5-Omni: Streaming Model (with Audio+Video Modalities)

Diffusion Transformer (DiT) + vocoder
(codec → mel-spectrogram) (mel → waveform)

Talker's own audio+text tokens so far
(for autoregressive modeling)

Embeddings of the generated text tokens

Thinker's hidden states for the whole input
(including text, vision, audio)



Similar to Qwen2.5-VL but with audio

[Qwen2.5-Omni Technical Report, 2025]

Different Strategies for Coherent Multimodal Generation…

	what is shared	what is separated
Chameleon	Everything: one vocabulary, one head, one loss	—
Janus	the transformer	the encoders and the heads
Transfusion	the transformer	the generative process: loss *and* sampling
Qwen-Omni	the context	the generator: a separate decoder for speech

… but coherence is assumed, never measured