

Unified-IO 2: Scaling Autoregressive Multimodal Models with Vision, Language, Audio, and Action

**Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya
Khosla, Ryan Marten, Derek Hoiem, Aniruddha Kembhavi
CVPR 2024**

Aarushi Wagh, Merna Bibars, Tenzin

Outline

- Problem Statement
- Related Works
- Approach
- Experiments & Results
- Limitations, Societal Implications
- Summary of Strengths, Weaknesses, Relationship to Other Papers

Problem Statement

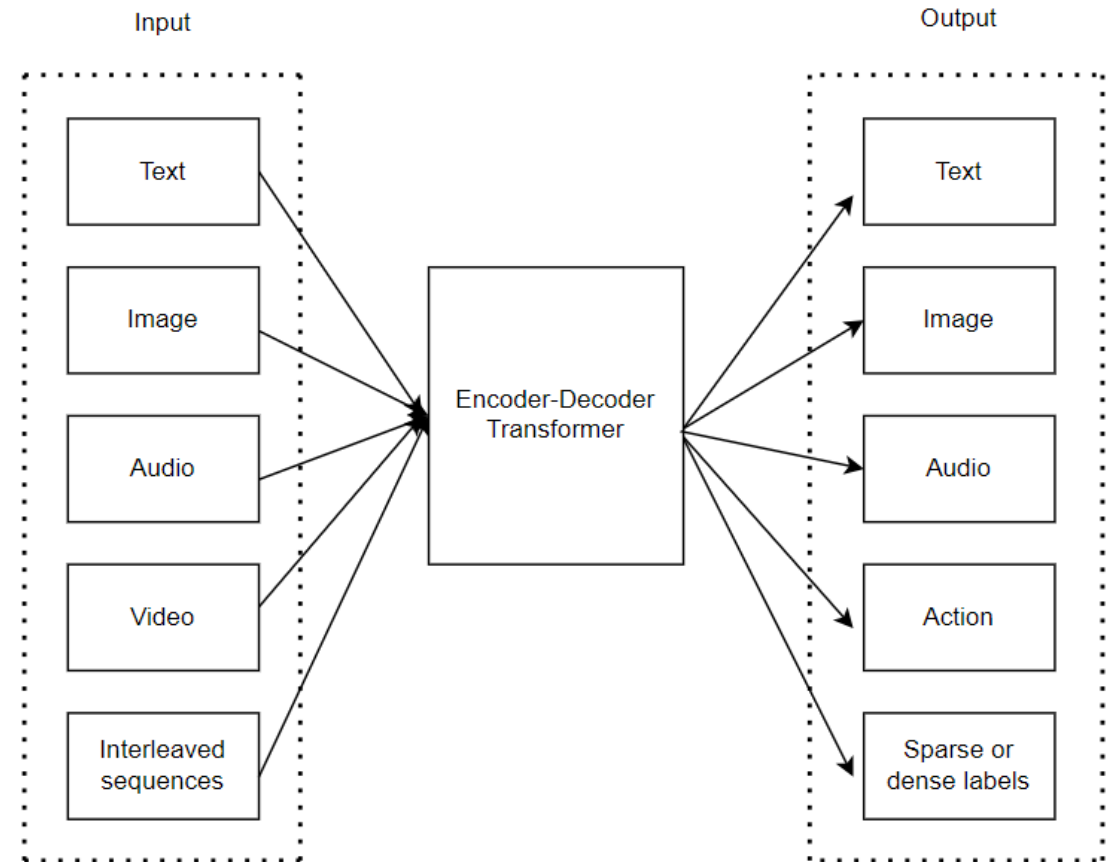
Problem Statement

Goals

- Unify multiple modalities - text, images, audio, video, and actions
- Address the challenges of training such a diverse multimodal model from scratch

Contributions

- First autoregressive model trained from scratch that can process and generate multiple modalities



Broad overview of Unified IO 2

Problem Statement












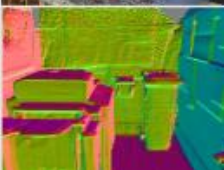
Contributions

- Pre-training
 - 1 billion image-text pairs, 1 trillion text tokens, 180 million video clips, 130 million interleaved image & text, 3 million 3D assets, and 1 million agent trajectories
- Instruction Tuning
 - Combined more than 120 datasets covering 220 tasks
- Architectural innovations
 - Multimodal mixture of denoiser objectives
 - Dynamic packing
- Stabilized training
 - 2D rotary embeddings
 - QK normalization
 - Scaled cosine attention mechanisms on the perceiver resampler

Problem Statement

Contributions

- Sets the new SOTA on the GRIT benchmark
- Matched or outperformed recently proposed VLMs
- Outperformed the closest competitor in image generation
- Capable of following free-form instructions, even those unseen during training

Task	Input Image	Input Query / Options	Output
Categorization		<i>[open_images_categories]</i>	<i>drill</i>
Localization		<i>kitchen & dining room table</i>	
Visual Question Answering		<i>Does this sofa have armrests?</i>	<i>yes</i>
Referring Expressions		<i>man on end black suit</i>	
Segmentation		<i>dolphin</i>	
Pose Keypoints		<i>person</i>	
Surface Normals			

GRIT Benchmark

Related Works

Related Works

1. MiniGPT-v2, InstructBLIP - Leveraged pre-trained LLMs and incorporated a visual encoder, and some extra visual instruction training to incorporate zero shot capabilities
2. Qwen-VL (OCR, image-text retrieval), M3IT (multilingual instruction tuning) - Added new functionalities to previous approach
3. PandaGPT, ImageBind-LLM, ChatBridge – Added other modalities
4. Multimodal generation:
 1. UNIFIED-IO, LaVIT, OFA, Emu and CM3Leon - generate tokens that a VQ-GAN can then decode into an image
 2. GILL, Kosmos-G and SEED - generate features that a diffusion model can use
 3. JAM - fuses pre-trained language and image generation models

CODI: Any-to-Any Generation via Composable Diffusion

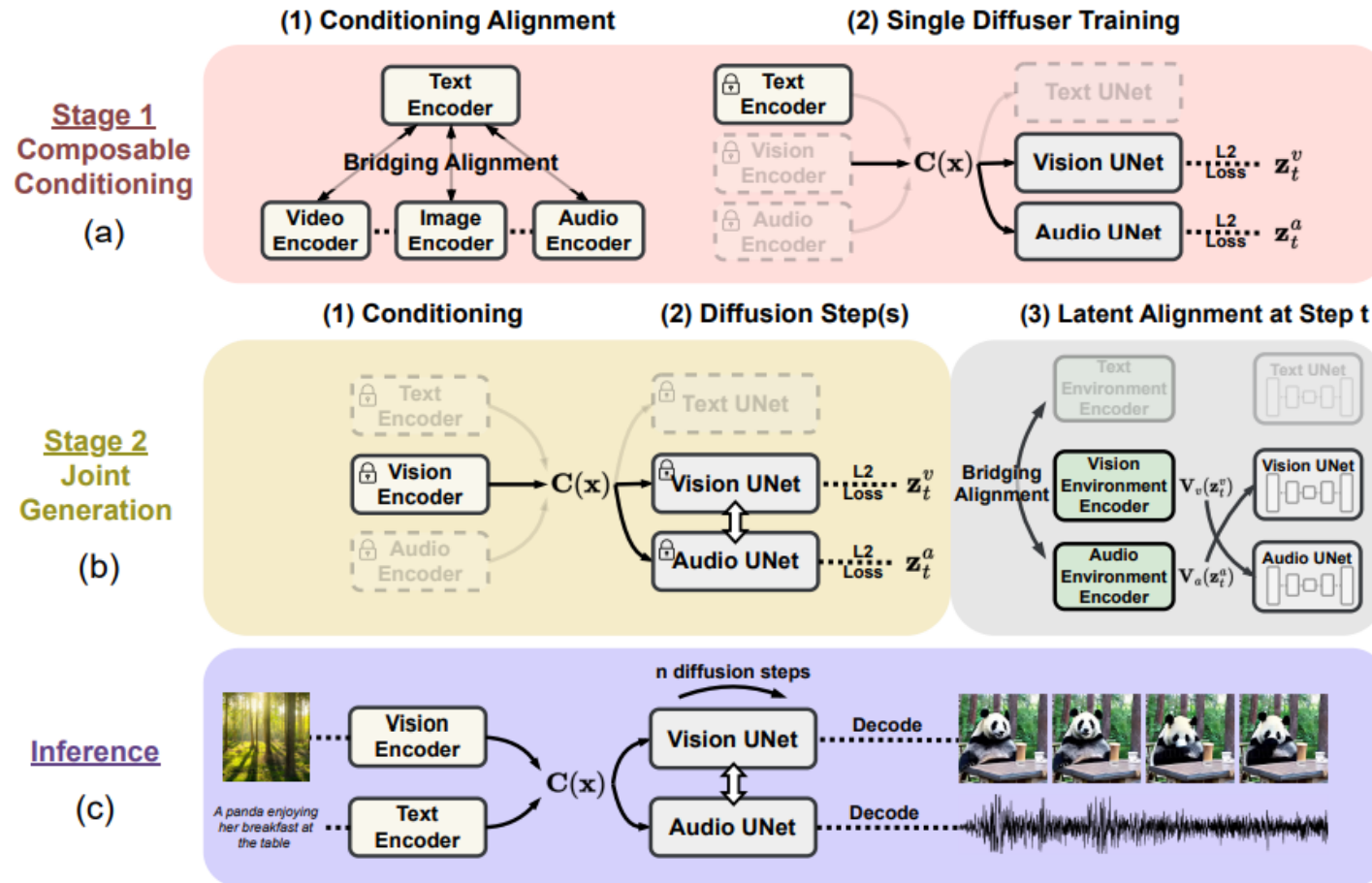


Figure 2: CoDi model architecture: (a) We first train individual diffusion model with aligned prompt encoder by “Bridging Alignment”; (b) Diffusion models learn to attend with each other via “Latent Alignment”; (c) CoDi achieves any-to-any generation with a linear number of training objectives.

CODI: Any-to-Any Generation via Composable Diffusion



Figure 3: Single-to-single modality generation. Clockwise from top left: text→image, image→text, image→video, audio→image.

CODI: Any-to-Any Generation via Composable Diffusion

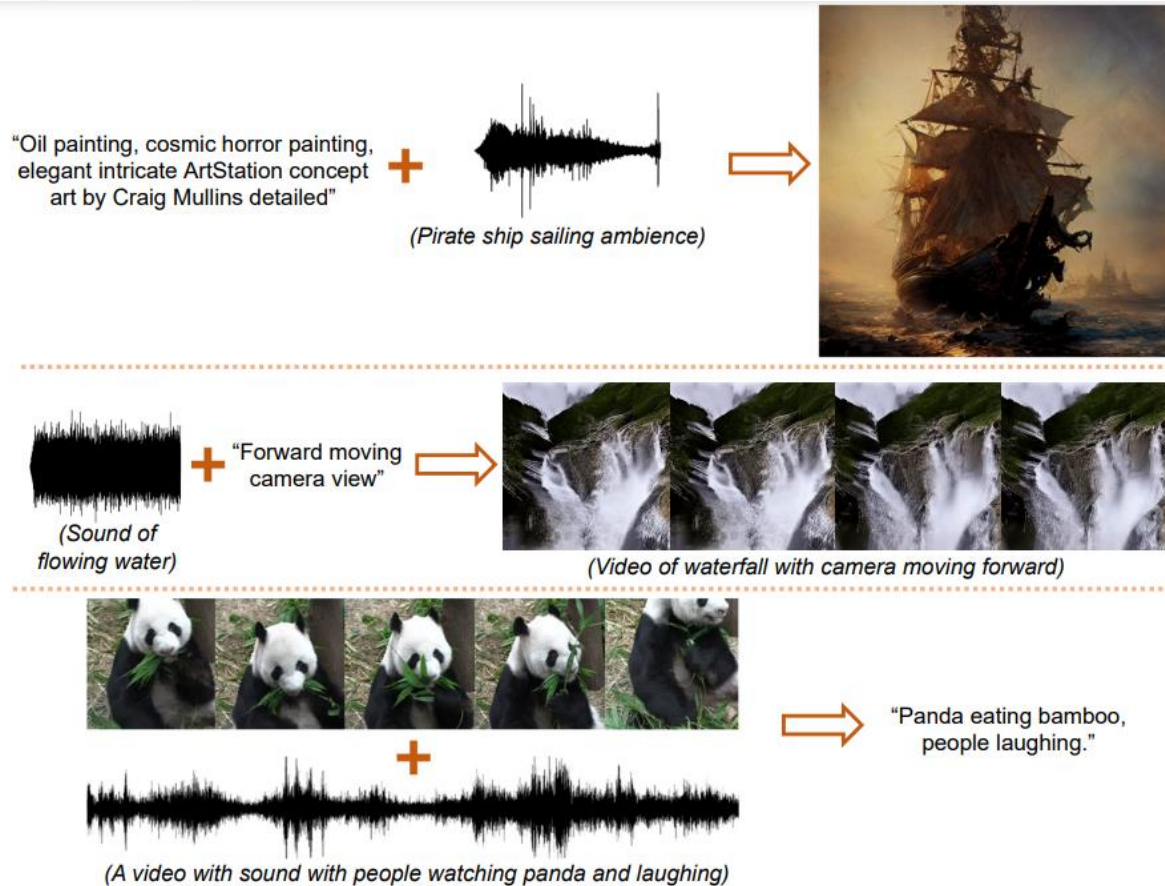


Figure 4: Generation with multiple input modality conditions. Top to bottom: text+audio→image, text+audio→video, video+audio→text.

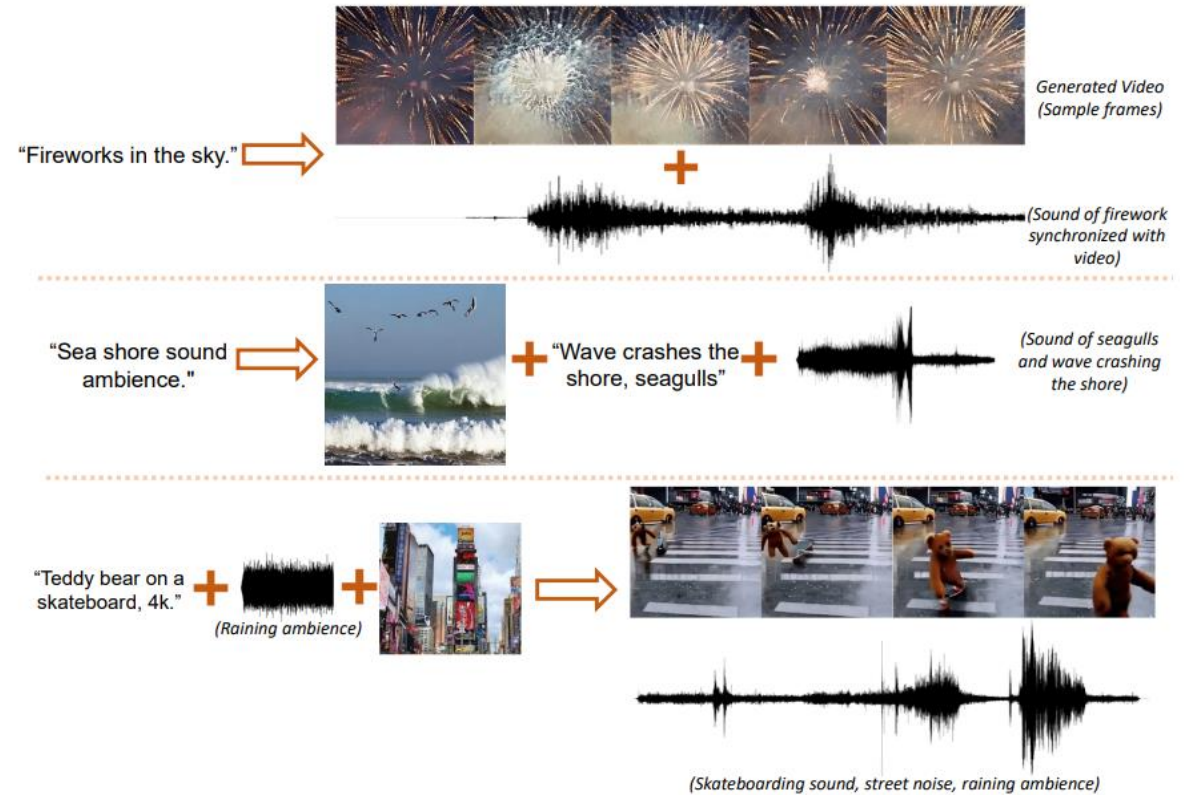


Figure 5: Joint generation of multiple output modalities by CoDi. From top to bottom: text→video+audio, text→image+text+audio, text+audio+image→video+audio.

Unified-IO: A Unified Model for Vision, Language, and Multi-Modal Tasks

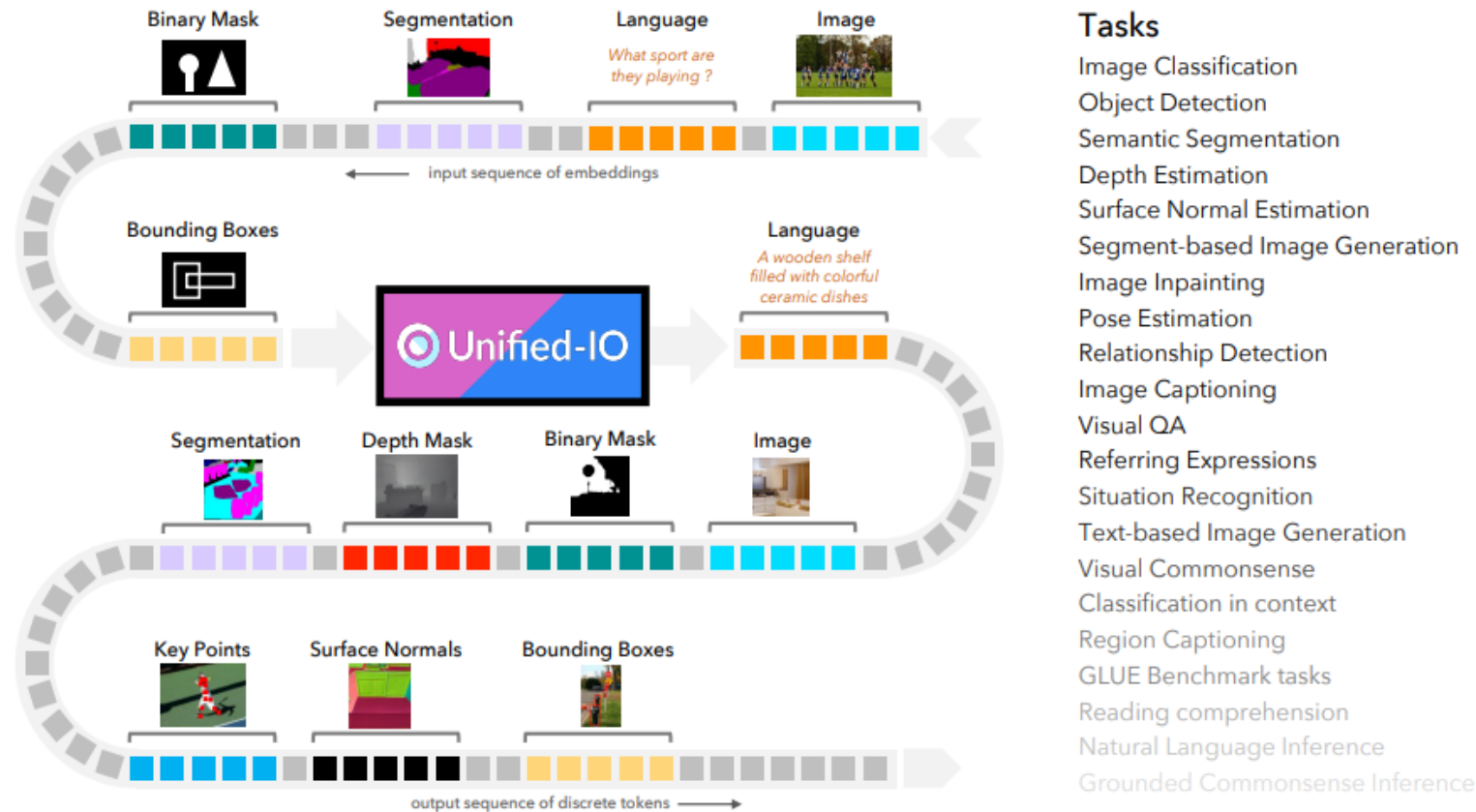
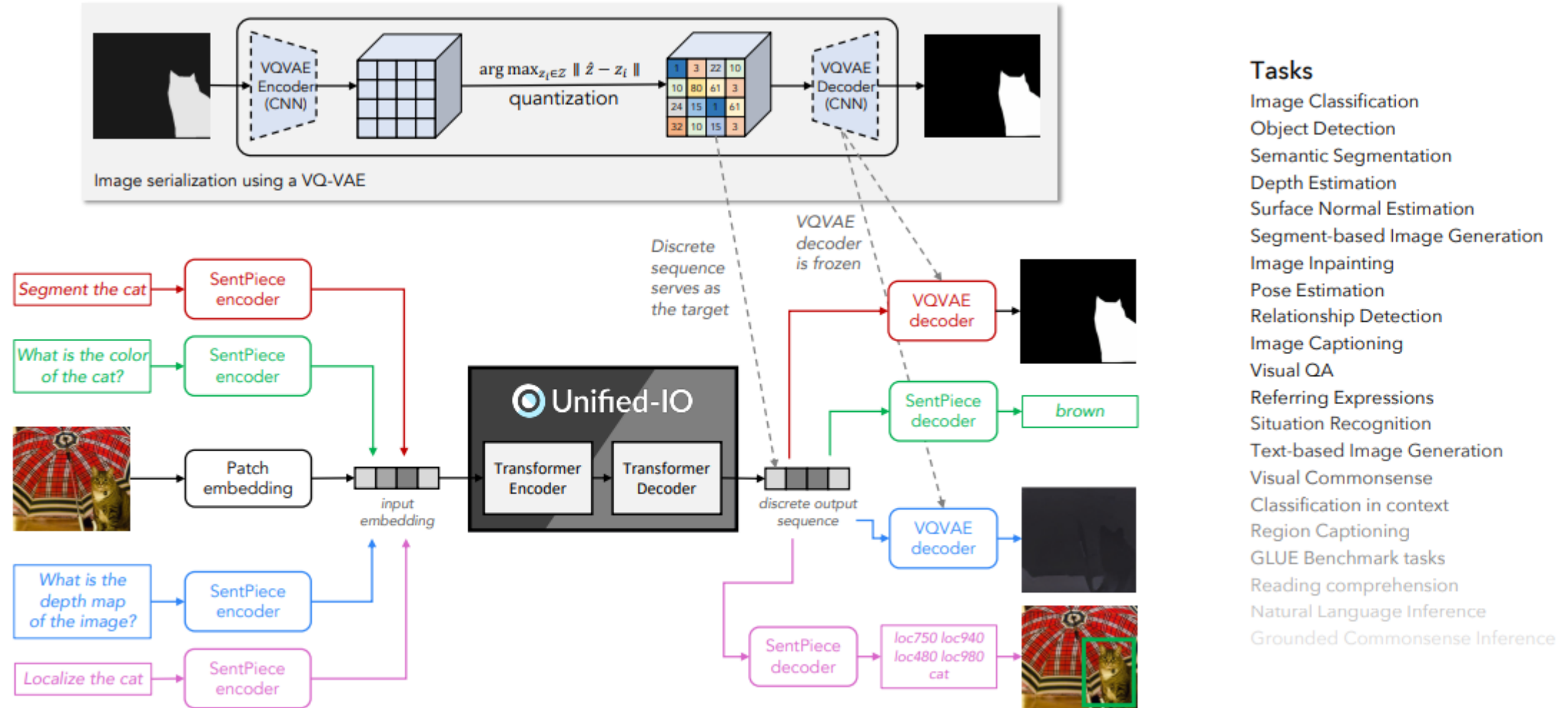


Figure 1: UNIFIED-IO is a single sequence-to-sequence model that performs a variety of tasks in computer vision and NLP using a unified architecture without a need for either task or modality-specific branches. This broad unification is achieved by homogenizing every task's input and output into a sequence of discrete vocabulary tokens. UNIFIED-IO supports modalities as diverse as images, masks, keypoints, boxes, and text, and tasks as varied as depth estimation, inpainting, semantic segmentation, captioning, and reading comprehension.

Unified-IO: A Unified Model for Vision, Language, and Multi-Modal Tasks



Tasks

- Image Classification
- Object Detection
- Semantic Segmentation
- Depth Estimation
- Surface Normal Estimation
- Segment-based Image Generation
- Image Inpainting
- Pose Estimation
- Relationship Detection
- Image Captioning
- Visual QA
- Referring Expressions
- Situation Recognition
- Text-based Image Generation
- Visual Commonsense
- Classification in context
- Region Captioning
- GLUE Benchmark tasks
- Reading comprehension
- Natural Language Inference
- Grounded Commonsense Inference

Figure 2: **Unified-IO**. A schematic of the model with four demonstrative tasks: object segmentation, visual question answering, depth estimation and object localization.

Approach

Architecture

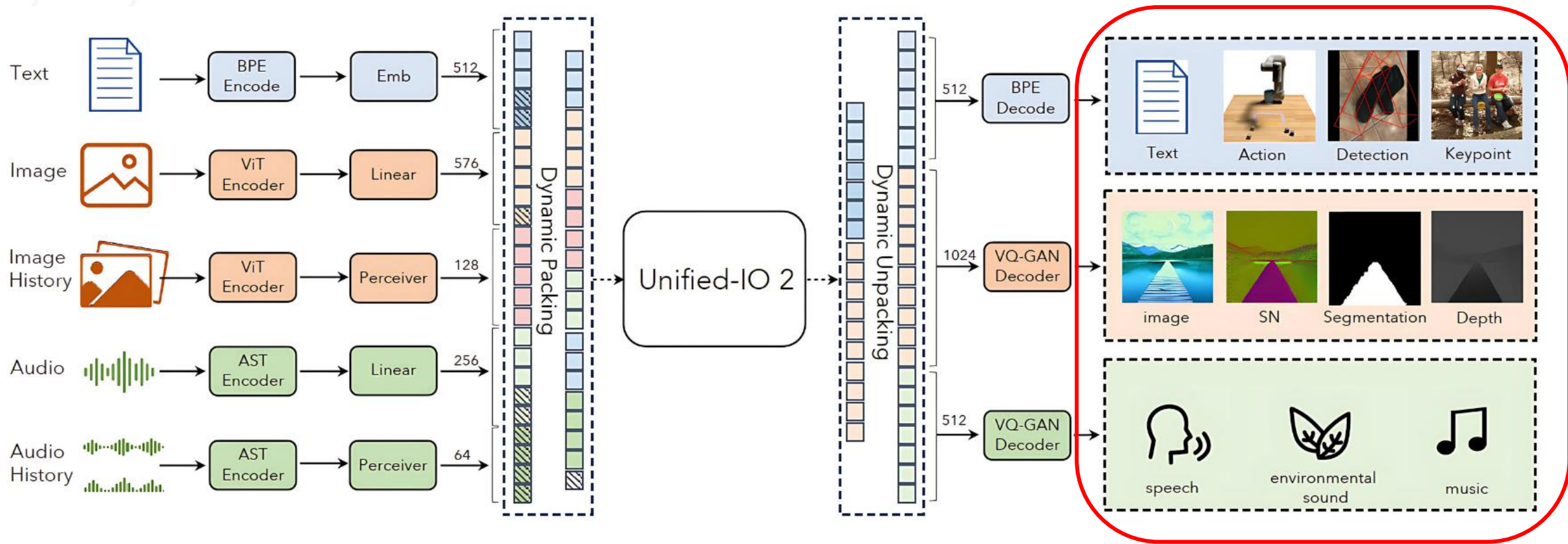


Figure 2. UNIFIED-IO 2 architecture. Input text, images, audio, or image/audio history are encoded into sequences of embeddings which are concatenated and used as input to an encoder-decoder transformer model. The transformer outputs discrete tokens that can be decoded into text, an image, or an audio clip.

Image Editing



Remove the dock



Paint this image like Van Gogh



Render a sunset

Reference Image Generation



Generate an image of a car with the model in the first image and the color in the second image.



Image Generation

Generate an image of an astronaut riding a horse in the forest. There is a river in front of them with water lilies.



Generate an image of an elephant swimming underwater. Aesthetic. Fantasy.



Multiview Image Completion



Add the missing details to the masked image (left) using the reference image (right).



Free Form VQA



What food could you make with these ingredients? Give me the recipe.

One delicious recipe using these ingredients is chocolate pudding! Here's the recipe:

Ingredients:
- 1 cup all-purpose flour, - 1/2 cup sugar
Instructions:
1. In a large bowl whisk together the flour, sugar...
2. In a separate bowl, mix together the eggs ...

Visual Parsing & Segmentation



Which fruits are in this image? List them in json format with the name of the fruit as the key and the color of the fruit as the value.

```
{  
  "banana": "Yellow",  
  "apple": "Red",  
  "grapes": "Green"  
}
```



Segment grapes and one apple

Depth & Surface Normal



Generate a depth image



Generate a surface normal map



Generate a depth & surface normal map

Keypoint Estimation



Find the visible keypoints corresponding to the person located in the highlighted region.



Tasks

Representation?

Audio Visual



Generate music about this scene.



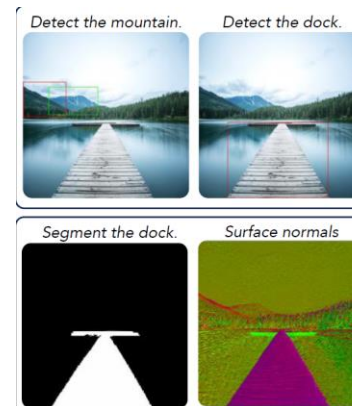
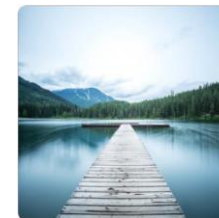
Locate the bounding boxes of the sound sources



3D Object Detection



2D Detection - Segmentation



3D Synthesis from Image

Input



Prompt

Complete a new image of the image following the implementation of the camera transformation θ $\sin(\phi)$ $\cos(\phi)$ r : ($\langle \text{extra_id_616} \rangle$ $\langle \text{extra_id_1011} \rangle$ $\langle \text{extra_id_309} \rangle$ $\langle \text{extra_id_729} \rangle$).

Prediction



GT



Robot Actions



Image Input

Text Input

[Image] [S] Given the initial observation $\langle \text{image_input} \rangle$ and prompt "Open the drawer.", predict the goal image.



Image Target



Modality Representation

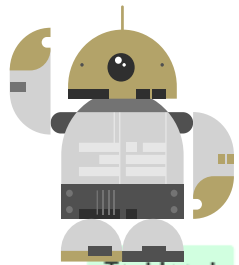
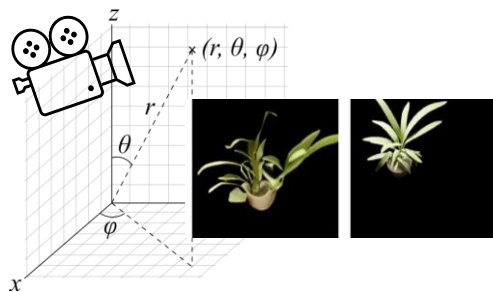
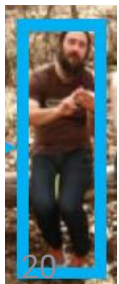
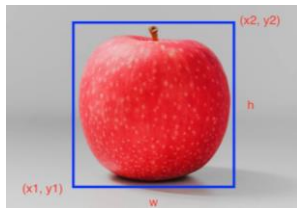
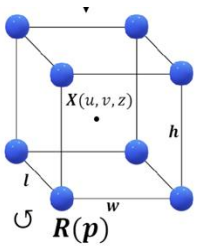
Text LLaMA BPE

Sentence: "It is raining."

Sub-word level tokenization

It is rain ing .

Continuous values



Text Input

[Text] [S] Imagine you are the robot in a silver pot to the right side of the table. and current observation image_input, guess the following action history.
 Visual Observation Δ Pos X Δ Pos Y Δ Pos Z Δ Rot X Δ Rot Y Δ Rot Z Gripper
 <image_history_1>: (<extra_id_696> <extra_id_705> <extra_id_726> <extra_id_767> <extra_id_683> <extra_id_711> <extra_id_200>)

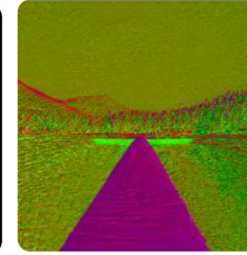
Images



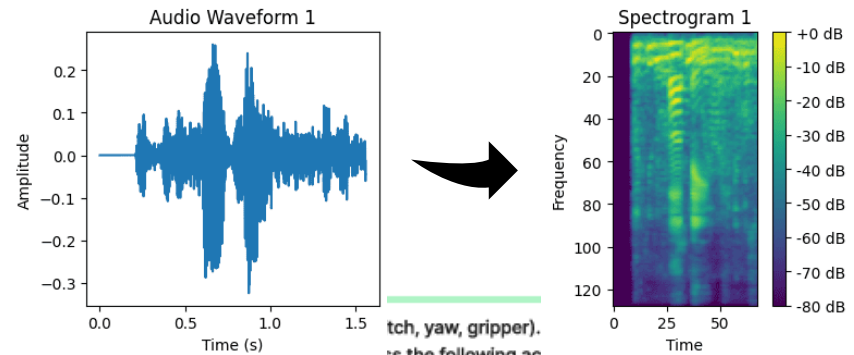
Segment the dock.



Surface normals



Audio



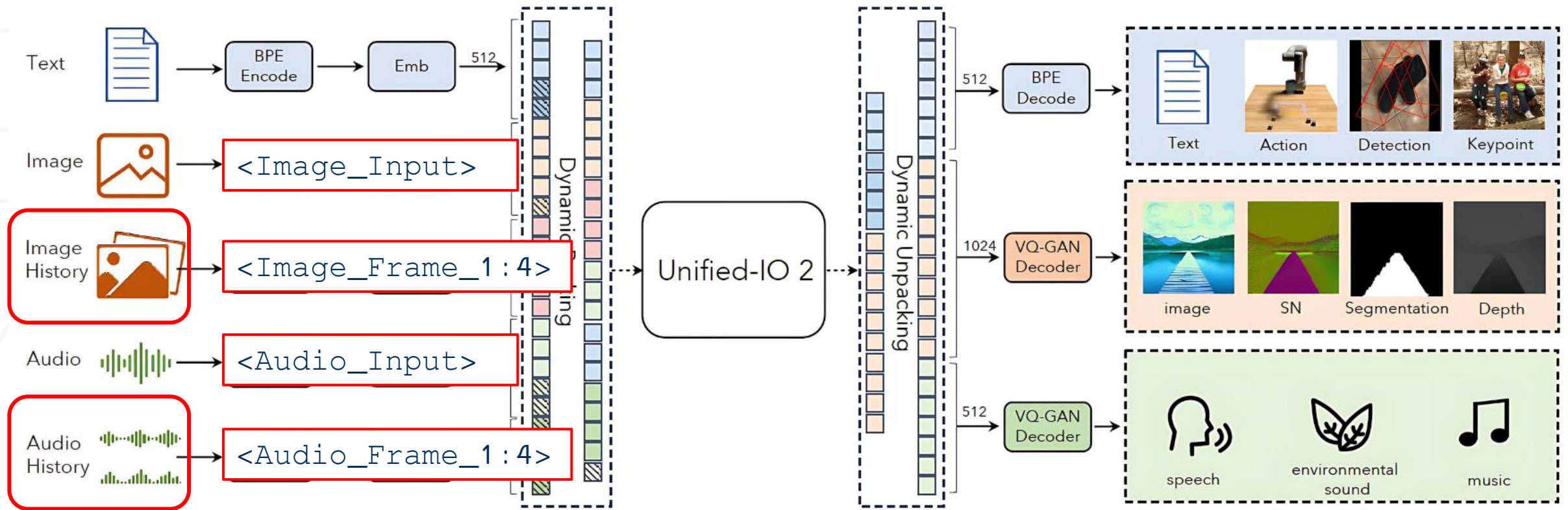


Figure 2. UNIFIED-IO 2 architecture. Input text, images, audio, or image/audio history are encoded into sequences of embeddings which are concatenated and used as input to an encoder-decoder transformer model. The transformer outputs discrete tokens that can be decoded into text, an image, or an audio clip.

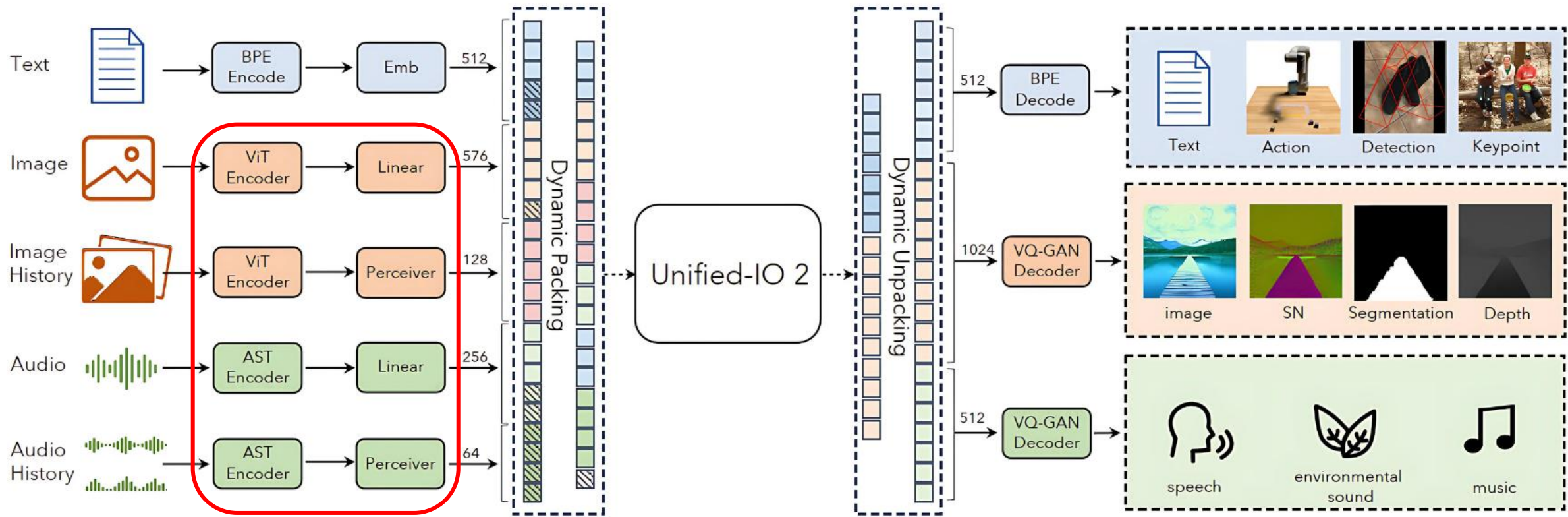
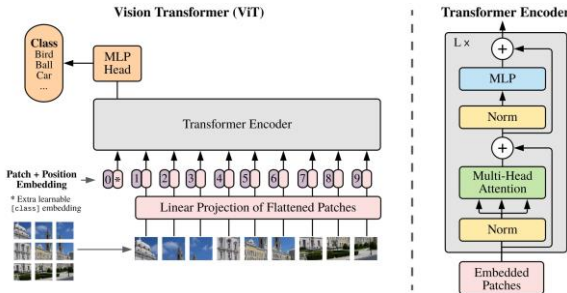


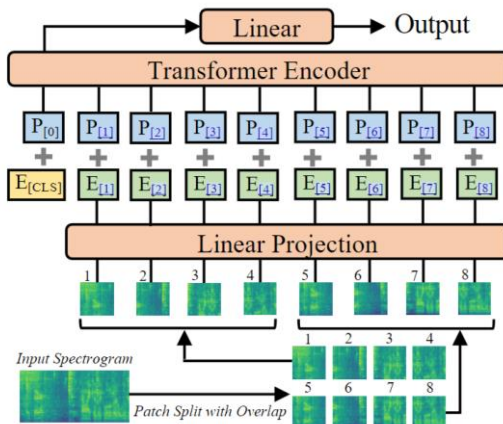
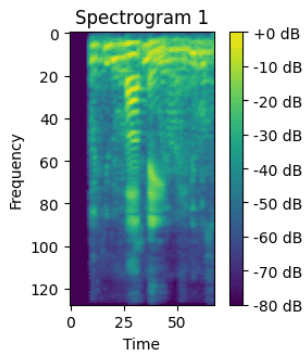
Figure 2. UNIFIED-IO 2 architecture. Input text, images, audio, or image/audio history are encoded into sequences of embeddings which are concatenated and used as input to an encoder-decoder transformer model. The transformer outputs discrete tokens that can be decoded into text, an image, or an audio clip.

Modality Encoding

Images



Audio



AST



ViT

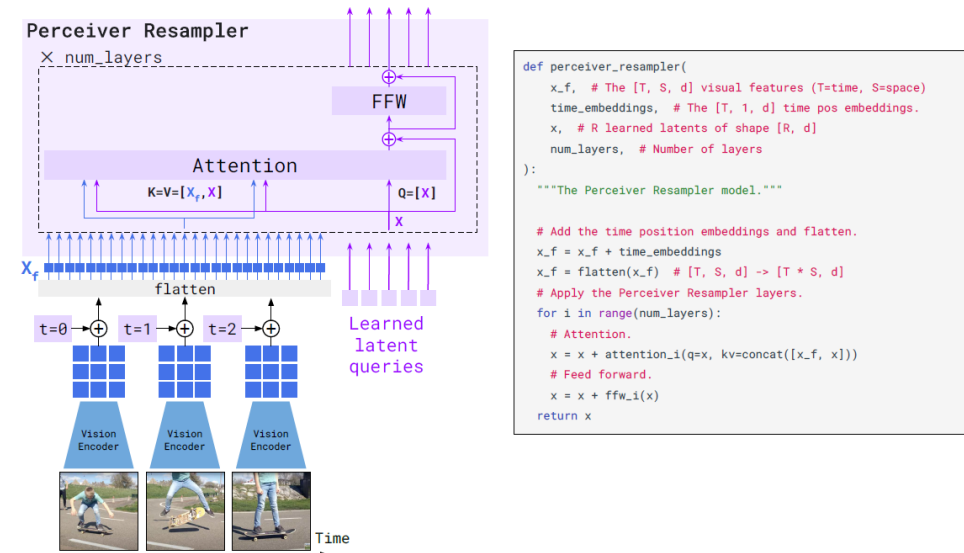


Figure 5: **The Perceiver Resampler** module maps a *variable* size grid of spatio-temporal visual features output by the Vision Encoder to a *fixed* number of output tokens (five in the figure), independently from the input image resolution or the number of input video frames. This transformer has a set of learned latent vectors as queries, and the keys and values are a concatenation of the spatio-temporal visual features with the learned latent vectors.

Perceiver Resampler



Architecture

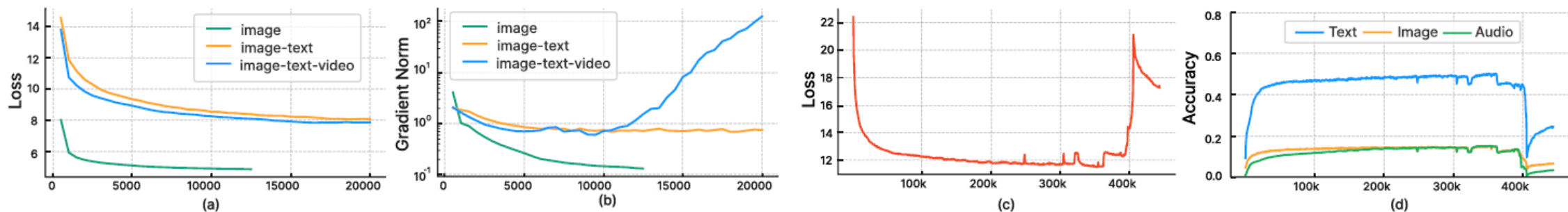
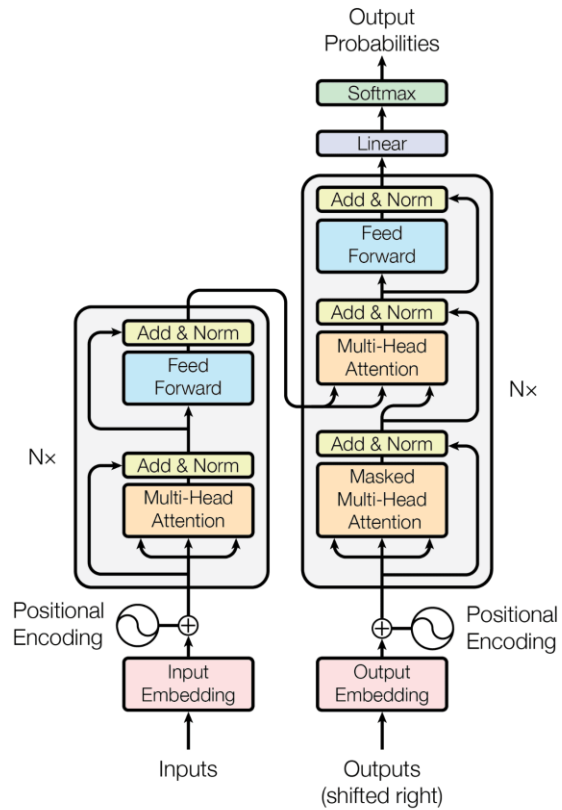


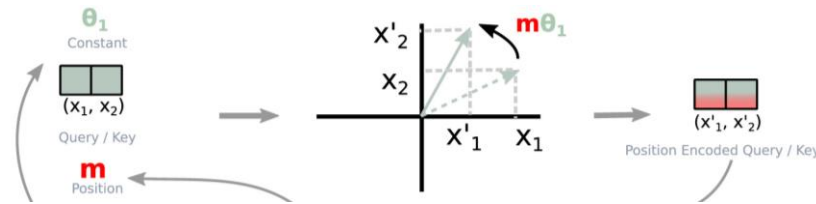
Figure 3. **Left:** Training loss (a) and gradient norms (b) on different modality mixtures. **Right:** Training loss (c) and next token prediction accuracy (d) of UIO-2_{XXL} on all modalities. Results were obtained before applying the proposed architectural improvements.

Architecture

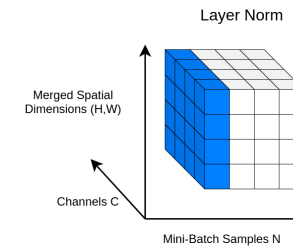


Improvements

(1) RoPE: 2D Rotary Positional Embeddings



(2) QK Normalization: LayerNorm



(3) Scaled Cosine Attention: Perceiver Resampler

$$\begin{array}{l}
 \mathbf{Q} \rightarrow \\
 \mathbf{K} \rightarrow \\
 \mathbf{V} \rightarrow
 \end{array}
 \left\{ \begin{array}{l}
 \text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) \\
 = \text{softmax} \left(\frac{\cos(\mathbf{Q}, \mathbf{K}) / \tau}{\sqrt{d_k}} \right) \mathbf{V}
 \end{array} \right.$$

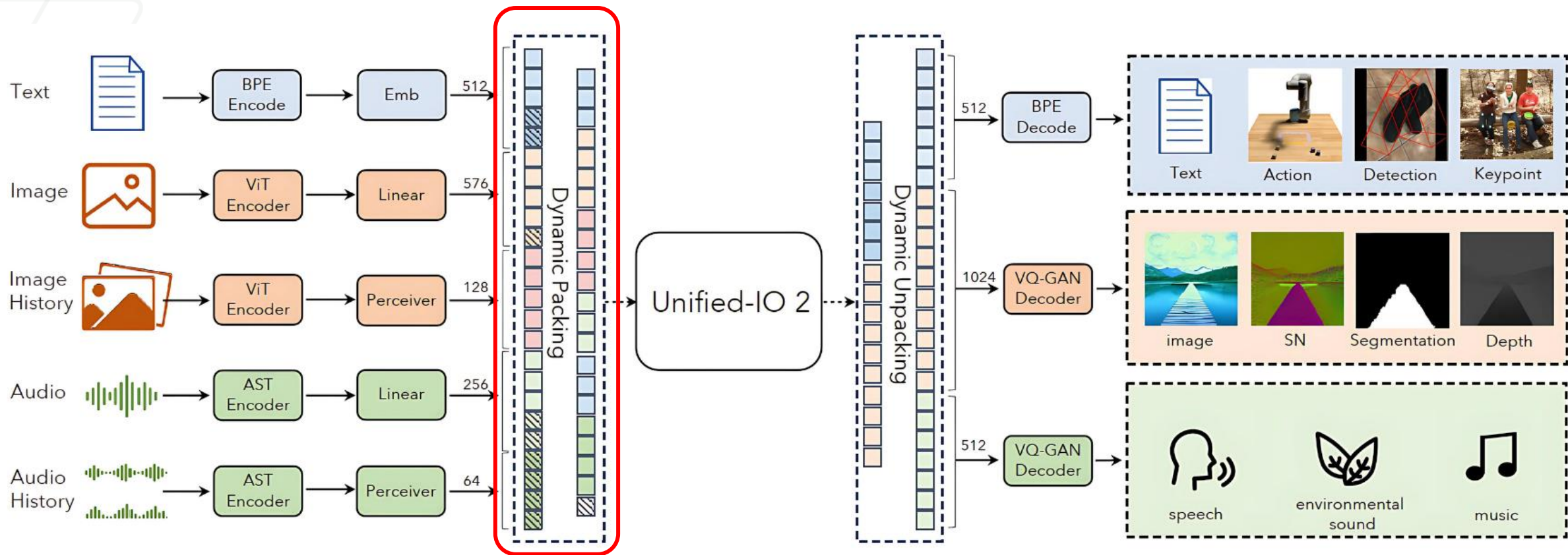


Figure 2. UNIFIED-IO 2 architecture. Input text, images, audio, or image/audio history are encoded into sequences of embeddings which are concatenated and used as input to an encoder-decoder transformer model. The transformer outputs discrete tokens that can be decoded into text, an image, or an audio clip.

Dynamic Packing

- Tokens of multiple examples are packed into a single sequence, and the attentions are masked to prevent the transformer from cross-attending between examples
- **4x training throughput**



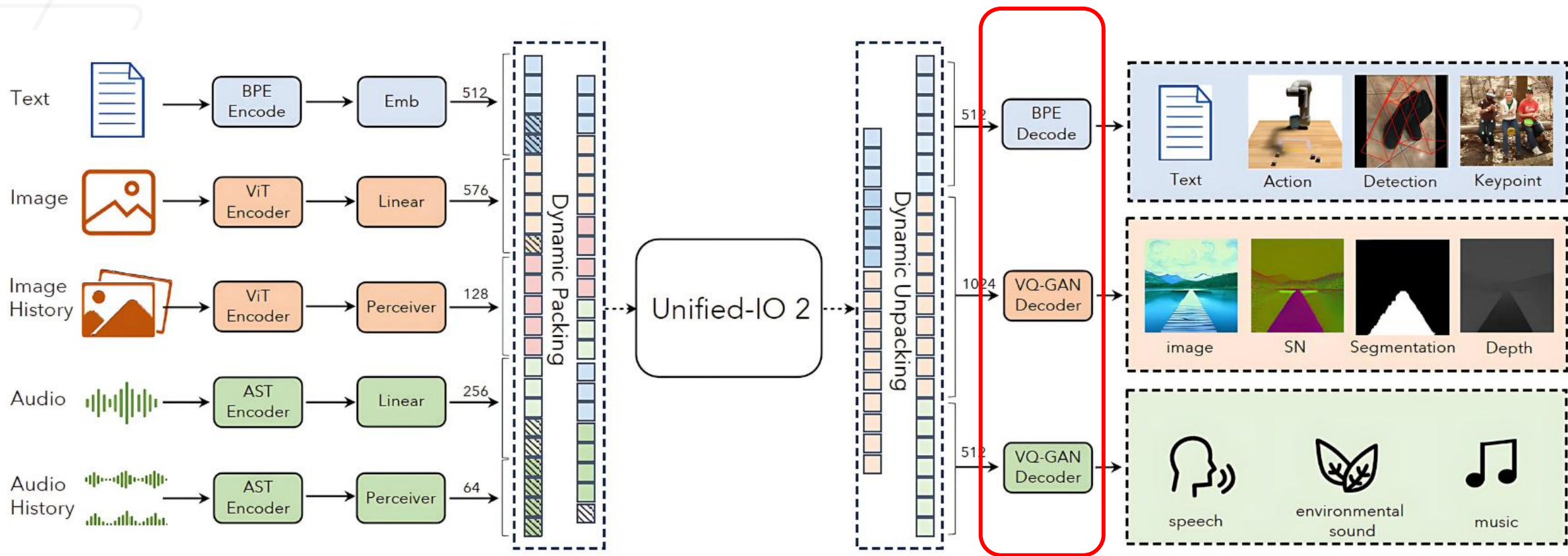
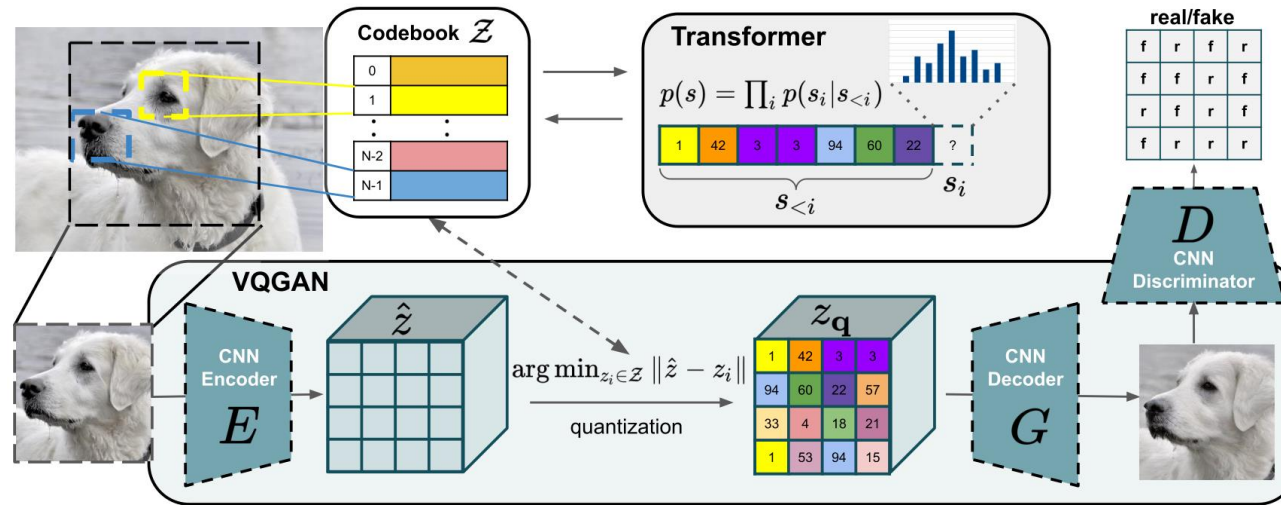


Figure 2. UNIFIED-IO 2 architecture. Input text, images, audio, or image/audio history are encoded into sequences of embeddings which are concatenated and used as input to an encoder-decoder transformer model. The transformer outputs discrete tokens that can be decoded into text, an image, or an audio clip.

Decoding

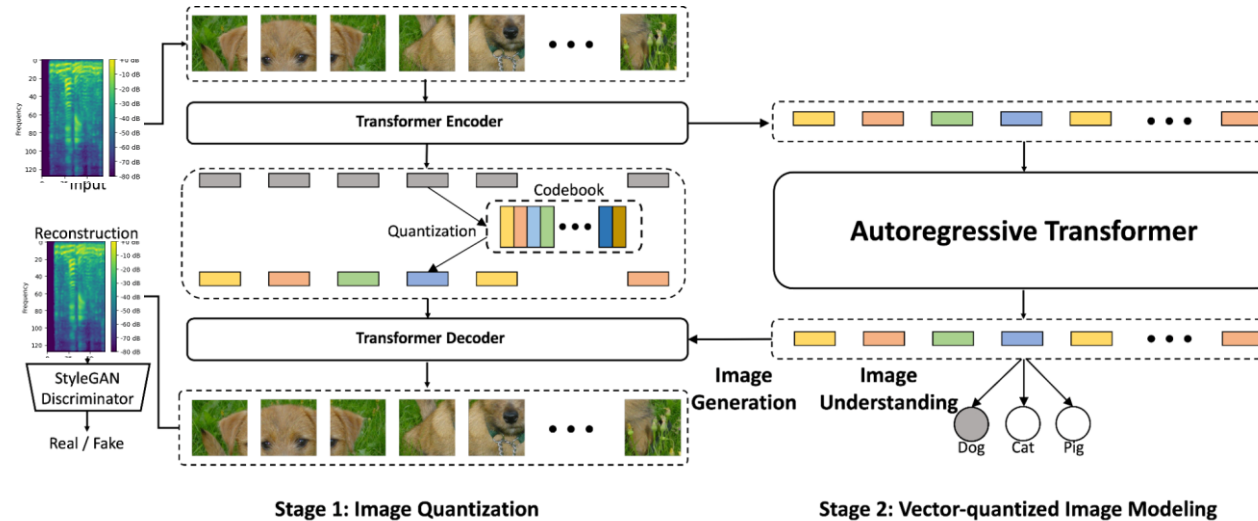
Images

VQ-GAN



Audio

ViT VQ-GAN



Model

Model	model dims	mlp dims	encoder lyr	decoder lyr	heads	Params
UIO-2 _L	1024	2816	24	24	16	1.1B
UIO-2 _{XL}	2048	5120	24	24	16	3.2B
UIO-2 _{XXL}	3072	8192	24	24	24	6.8B

Table 1. Size variant of UNIFIED-IO 2.

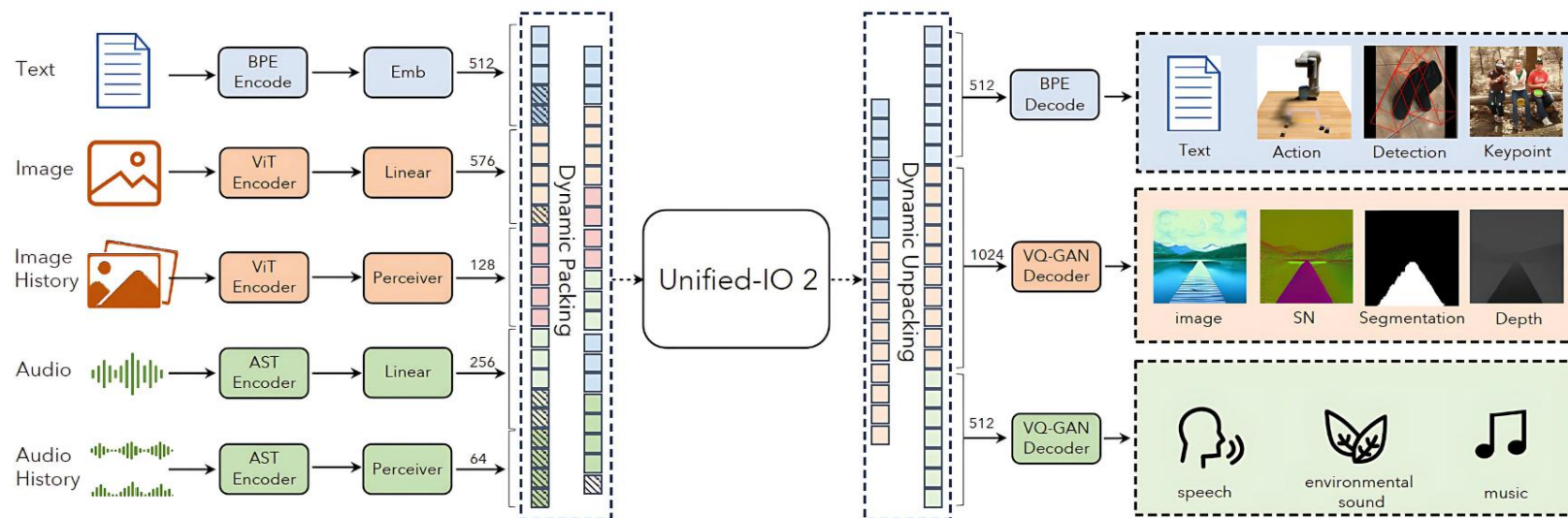


Figure 2. UNIFIED-IO 2 architecture. Input text, images, audio, or image/audio history are encoded into sequences of embeddings which are concatenated and used as input to an encoder-decoder transformer model. The transformer outputs discrete tokens that can be decoded into text, an image, or an audio clip.

Training Objectives

Multimodal Mixture of Denoisers

Text

UL2

[R] masked language modeling (MLM)

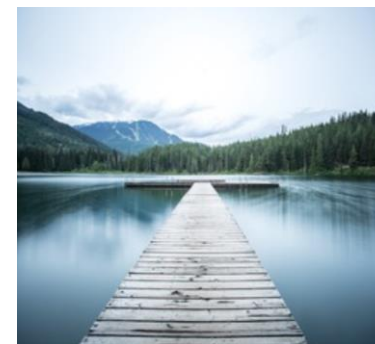
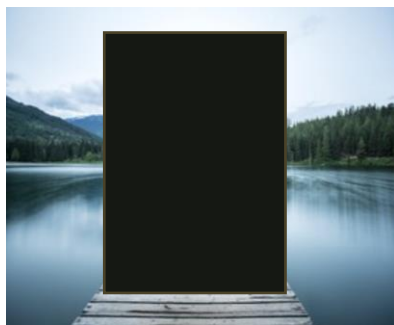
[S] causal language modeling

[X] extreme span corruption

Multimodal Mixture of Denoisers

Image & Audio

[R] \rightarrow x%
masked



[S]
causal modality



Generate music about this scene.



Multimodal Mixture of Denoisers

Image & Audio

[R] \rightarrow x%
masked

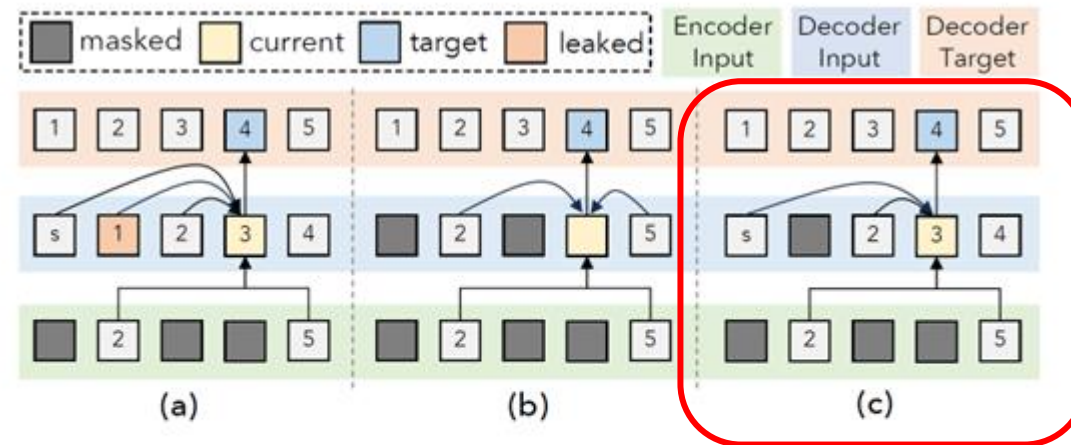


Figure 5. Different training paradigms in masked image modeling: (a) autoregressive, (b) mask auto-encoder, (c) autoregressive with dynamic masking. Our proposed paradigms can maintain causal generation while avoiding information leaks in the decoder.

Pre-Training Objective

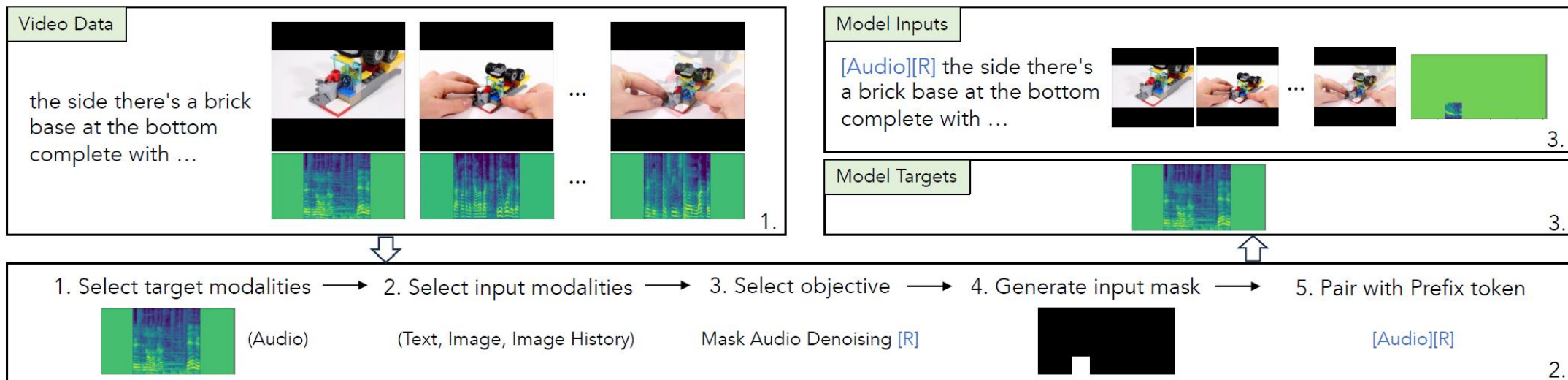


Figure 7. Construction of training samples from video data for the model's input and target. Given the video, we first extract the video frames and the corresponding audio spectrograms and transcript. Then, the data pass through a random selection process to determine the target modality, input modalities, training objective, input mask *etc.* The model's final input and target are shown in the top right.

Datasets

Pre-Training Data

- Total size: **8.5 Billion**
- **1 billion** image-text pairs, **6.6 billion** text, **180 million** video clips, **130 million** interleaved image & text, **3 million** 3D assets, and **1 million** agent trajectories

	Size	Rate	Text	Sparse	Dense	Image	Audio	ImageH	AudioH	Text	Sparse	Dense	Image	Audio
Text	6.6b	33.0	✓	-	-	-	-	-	-	✓	-	-	-	-
MC4 [201]	5.0b	11.7	✓	-	-	-	-	-	-	✓	-	-	-	-
C4 [68]	266m	10.6	✓	-	-	-	-	-	-	✓	-	-	-	-
Stack [95]	147m	3.55	✓	-	-	-	-	-	-	✓	-	-	-	-
RedPajama CC [32]	1.2b	3.55	✓	-	-	-	-	-	-	✓	-	-	-	-
Wikipedia	6.8m	1.42	✓	-	-	-	-	-	-	✓	-	-	-	-
RedPajama Book [32]	13m	1.06	✓	-	-	-	-	-	-	✓	-	-	-	-
Stack-Markdown [95]	34m	1.06	✓	-	-	-	-	-	-	✓	-	-	-	-
Image/Text	970m	31.3	✓	-	-	✓	-	-	-	✓	-	-	✓	-
LAION Aesthetics v2.5 [158]	491m	17.7	✓	-	-	✓	-	-	-	-	-	-	✓	-
LAION-400M [159]	346m	8.95	✓	-	-	✓	-	-	-	✓	-	-	-	-
CC12M [23]	11m	1.48	✓	-	-	✓	-	-	-	✓	-	-	✓	-
RedCaps [42]	12m	1.39	✓	-	-	✓	-	-	-	✓	-	-	✓	-
Web Images	107m	1.33	✓	-	-	✓	-	-	-	✓	-	-	✓	-
CC3M [163]	3.0m	0.49	✓	-	-	✓	-	-	-	✓	-	-	✓	-

Contd... in supplementary materials

Instruction Tuning Data

- **220 tasks** from over **120 datasets**
- **60% prompting data**, meaning supervised datasets combined with prompts
- Catastrophic forgetting → **30%** of the data is carried over from **pre-training**
- **6%** task augmentation data we build by constructing novel tasks using existing data sources, which enhances existing tasks and increases task diversity.
- The remaining **4%** consists of **free-form text** to enable chat-like responses.
 - Pre-training BS = 512, 1.5 M steps
 - Instruction-tuning BS = 256, 1.5 M steps

	Size	Rate	Datasets
Image Generation	506m	17.6	21
Image from Text	497m	10.6	5
Controllable Image Editing	3.0m	2.92	4
Image Editing	1.1m	1.66	3
Next Frame Generation	24k	0.96	2
Image Inpainting	1.0m	0.79	3
View Synthesis	4.2m	0.60	4
Audio Generation	164m	7.50	9
Audio from Text	19m	5.62	8
Audio from Video	145m	1.88	1
Image Understanding	53m	17.8	73
VQA	5.8m	6.23	31
Image Captioning	32m	4.25	14
Region Classification	6.1m	2.41	4
Image Tagging	3.8m	2.38	8
Relationship Prediction	0.8m	1.41	6
Region Captioning	3.5m	0.60	1
Image Instruction Following	0.4m	0.37	6
Image Pair QA	0.1m	0.17	3

Contd... in supplementary materials

Experiments Results

Experiments and Results

Experiments -> too complicated and lots of numbers and evals.

We will adopt a precise thinking framework. I call -> **Pierce**.

We will ride on **Pierce** to weed through the noise in all evals, one by one, as we collect signal that stays with us till we reach the end.

Pierce Framework:

1. Describe the eval.
2. Keep an eye for numbers that jump out -> Where are the **bolds** ?
3. What's the diff b/w the best and 2nd best ?
 - "Unmotivated" best reasoning explaining the diff.
 - Compare approaches based on:
 - Model params
 - Data size
 - Approach Diff (if necessary)
 - Data Type (if necessary)
4. Collect Signal -> Trends that stand the test of ALL evals "without" inconsistencies.

Zero Shot Analysis – Not much Signal, but a good start.

Common Sense NLI | Faithfulness of text-2-image | spatio-temporal comprehension | captioning audios

Data
1 Trillion tokens
1 Trillion tokens
LAION-400M (Under 1.5B params)
LAION-120M + MC4 60M

UIO-2 Data

Text -> 6.6 Billion Tokens

Image-text -> 970m pairs

Image/Text	970m	31.3
LAION Aesthetics v2.5 [158]	491m	17.7
LAION-400M [159]	346m	8.95
CC12M [23]	11m	1.48
RedCaps [42]	12m	1.39
Web Images	107m	1.33
CC3M [163]	3.0m	0.49

Method	HellaSwag↑	TIFA↑	SEED-S↑	SEED-T↑	AudioCaps↓
LLaMA-7B [177]	76.1	-	-	-	-
OpenLLaMa-3Bv2 [55]	52.1	-	-	-	-
SD v1.5 [154]	-	78.4	-	-	-
OpenFlamingo-7B [9]	-	-	34.5	33.1	-

UIO-2 _L	38.3	70.2	37.2	32.2	3.08
UIO-2 _{XL}	47.6	77.2	40.9	34.0	3.10
UIO-2 _{XXL}	54.3	78.7	40.7	35.0	3.02

GRIT -> Signal !! for Uni-Modal vs Multi-Modal

~ FrameNet+NYUV2

~1M images

~ 3.2M images

10.2M human-loop masks

		Categorization		Localization		VQA		Refexp		Segmentation		Keypoint		Normal		All	
		ablation	test	ablation	test	ablation	test	ablation	test	ablation	test	ablation	test	ablation	test	ablation	test
0	NLL-AngMF [11]	-	-	-	-	-	-	-	-	-	-	-	-	49.6	50.5	7.2	7.1
1	Mask R-CNN [70]	-	-	44.7	45.1	-	-	-	-	26.2	26.2	70.8	70.6	-	-	20.2	20.3
2	GPV-1 [65]	33.2	33.2	42.8	42.7	50.6	49.8	25.8	26.8	-	-	-	-	-	-	21.8	21.8
3	CLIP [146]	48.1	-	-	-	-	-	-	-	-	-	-	-	-	-	6.9	-
4	OFA _{LARGE} [186]	22.6	-	-	-	72.4	-	61.7	-	-	-	-	-	-	-	22.4	-
5	GPV-2 [89]	54.7	55.1	53.6	53.6	63.5	63.2	51.5	52.1	-	-	-	-	-	-	31.9	32.0
5	DINO + SAM [94, 139]	-	-	66.0	66.0	-	-	-	-	60.2	60.1	-	-	-	-	18.0	18.0
6	UNIFIED-IO _{SMALL}	42.6	-	50.4	-	52.9	-	51.1	-	40.7	-	46.5	-	33.5	-	45.4	-
7	UNIFIED-IO _{BASE}	53.1	-	59.7	-	63.0	-	68.3	-	49.3	-	60.2	-	37.5	-	55.9	-
8	UNIFIED-IO _{LARGE}	57.0	-	64.2	-	67.4	-	74.1	-	54.0	-	67.6	-	40.2	-	60.7	-
9	UNIFIED-IO _{XL}	61.7	60.8	67.0	67.1	74.5	74.5	78.6	78.9	56.3	56.5	68.1	67.7	45.0	44.3	64.5	64.3
9	UIO-2 _L	70.1	-	66.1	-	67.6	-	66.6	-	53.8	-	56.8	-	44.5	-	60.8	-
10	UIO-2 _{XL}	74.2	-	69.1	-	69.0	-	71.9	-	57.3	-	68.2	-	46.7	-	65.2	-
11	UIO-2 _{XXL}	74.9	75.2	70.3	70.2	71.3	71.1	75.5	75.5	58.2	58.8	72.8	73.2	45.2	44.7	66.9	67.0

Remains, Unclear:

Are the gains, because of

1. Modalities
2. Unification
3. Just more raw data.
4. Just more model params.

Contradiction -> How come, despite less data, and model params:

1. DINO+SAM is better in segmentation ?
2. NLL-AngMF is better in surface Normal tasks ?

Generation Comparisons – Signal Persists

Metrics:

1. FAD/FID -> similarity to ground truth
2. TIFA -> faithfulness in generation to prompt.
3. IS -> sample diversity and quality
4. KL -> prob. Dist. b/w target and sample features

Audio:

1. AudioLDM wins, plausible reasons?
 1. Diffusion models vs VQ-GAN
 2. Specific task fine-tuning
2. Contradictions?
 1. UIO2 data >> AudioLDM data (AudioSet mainly)
 2. Diversity (IS) should be better, coz of large-train set.

Action:

- VIMA models win reasons?
1. Specifically tuned for this task only. No audio/video distractions?
 2. Trick -> object centric features
- Contradictions?
1. Less than 300M params.
 2. 1M trajectories vs 650k trajectories
 3. Uses a "non". unified way and beats "unified" IO2
 1. Image I/p -> Mask-RCNN -> localized objects -> VIT -> o/p image tokens
 2. Image Tokens -> MLP -> t5 text-only base
 3. text I/p -T5 base

0.4B param + 30M + CLIP
1.45B param + LAION-400M

Method	Image		Audio			Action
	FID↓	TIFA↑	FAD↓	IS↑	KL↓	Succ.↑
minDALL-E [37]	-	79.4	-	-	-	-
SD-1.5 [154]	-	78.4	-	-	-	-
AudioLDM-L [117]	-	-	1.96	8.13	1.59	-
AudioGen [101]	-	-	3.13	-	2.09	-
DiffSound [203]	-	-	7.75	4.01	2.52	-
VIMA [87]	-	-	-	-	-	72.6
VIMA-IMG [87]	-	-	-	-	-	42.5
CoDi [174]	11.26	71.6	1.80	8.77	1.40	-
Emu [172]	11.66	65.5	-	-	-	-
UIO-2 _L	16.68	74.3	2.82	5.37	1.93	50.2
UIO-2 _{XL}	14.11	80.0	2.59	5.11	1.74	54.2
UIO-2 _{XXL}	13.39	81.3	2.64	5.89	1.80	56.3

CoDi vs UIO2 -> High Signal !!!! Do u see it??

CoDi Dataset

Categories	Tasks	Datasets	# of samples
Image + Text	Image→Text, Text→Image Text→Image+Text	Laion400M [42]	400M
Audio + Text	Text→Audio, Audio→Text, Text→Audio+Text, Audio-Text CT	AudioSet [16] AudioCaps [24] Freesound 500K BBC Sound Effect	900K* 46K 2.5M 30K
Audiovisual	Image→Audio, Image→Video+Audio	AudioSet SoundNet [3]	900K* 1.0M*
Video	Text→Video, Image→Video, Video-Text CT	Webvid10M [4] HD-Villa-100M [54]	10.7M 100M

UIO-2 Dataset

	Size	Rate	Datasets
Image Generation	506m	17.6	21
Image from Text	497m	10.6	5
Controllable Image Editing	3.0m	2.92	4
Image Editing	1.1m	1.66	3
Next Frame Generation	24k	0.96	2
Image Inpainting	1.0m	0.79	3
View Synthesis	4.2m	0.60	4
Audio Generation	164m	7.50	9
Audio from Text	19m	5.62	8
Audio from Video	145m	1.88	1

Video		
	181m	25.0
YT-Temporal [215]	146m	13.7
ACAV [105]	17m	3.98
HD-VILA [200]	7.1m	2.75
AudioSet [54]	1.7m	2.75
WebVid [13]	9.2m	1.23
Ego4D [60]	0.7m	0.55

Inferences:

CoDi beats UIO2 , plausible reasons

1. Diffusion Model. (better quality than GAN)

Contradiction

1. Codi's audio data is smaller than UIO2's.

2. Overall, less data, and same # of modalities.

UIO2 beats CoDi in faithfulness to prompts. What could this mean? Better text-image grounding, but quality of image is limited by VQ-GAN ?

Vision-Language Tasks – Signal Persists

Method	VQA ^{v2}	OKVQA	SQA	SQA ^I	Tally-QA	RefCOCO	RefCOCO+	RefCOCO-g	COCO-Cap.	POPE	SEED	MMB
InstructBLIP (8.2B)	-	-	-	79.5	68.2 [†]	-	-	-	102.2	-	53.4	36
Shikra (7.2B)	77.4	47.2	-	-	-	87.0	81.6	82.3	117.5	84.7	-	58.8
Ferret (7.2B)	-	-	-	-	-	87.5	80.8	83.9	-	85.8	-	-
Qwen-VL (9.6B)	78.8	58.6	-	67.1*	-	89.4	83.1	85.6	131.9	-	-	38.2
mPLUG-Owl2 (8.2B)	79.4	57.7	-	68.7*	-	-	-	-	137.3	86.2	57.8	64.5
LLaVa-1.5 (7.2B)	78.5	-	-	66.8*	-	-	-	-	-	85.9	58.6	64.3
LLaVa-1.5 (13B)	80.0	-	-	71.6*	72.4 [†]	-	-	-	-	85.9	61.6	67.7
Single Task SoTA	86.0 [29]	66.8 [77]	90.9 [119]	90.7 [34]	82.4 [77]	92.64 [202]	88.77 [187]	89.22 [187]	149.1 [29]	-	-	-
UIO-2 _L (1.1B)	75.3	50.2	81.6	78.6	69.1	84.1	71.7	79.0 [◇]	128.2	77.8	51.1	62.1
UIO-2 _{XL} (3.2B)	78.1	53.7	88.8	87.4	72.2	88.2	79.8	84.0 [◇]	130.3	87.2	60.2	68.1
UIO-2 _{XXL} (6.8B)	79.4	55.5	88.7	86.2	75.9	90.7	83.1	86.6 [◇]	125.4	87.7	61.8	71.5

Inferences:

1. Marginal improvements.
 - Except SQA1 (but there's a star XD)
2. Single Task >> Multi Task, However they are still close.

Single Task SOTA usually have something "additional" giving them the edge.

1. RefCOCO+/g : 1.5B Laion-2B images
2. QKVQA, SQA: Pali –VLM (5B params) + fine-tuned + reasoning paradigm

Table -> What's the meaning of Life? -> Meaningless

(*) - zero shot

(**) - few shot

UIO2 – Instruction Tuned.

Only 3/8 methods are fairly compared.

Only 6 / 10 datasets maybe fairly inspected. 2 of those are from flamingo-80B, so unfair again.

Only 4 remain, UIO2 loses in 3 of those.

Impossible to draw any inferences for majority of models/data.

For which it's possible -> UIO2 underperforms anyways.

Method	Video						Audio			
	Kinetics-400 [90]	VATEXCaption [190]	MSR-VTT [199]	MSRVTT-QA [198]	MSVD-QA [198]	STAR [196]	SEED-T [106]	VGG-Sound [24]	AudioCaps [93]	Kinetics-Sounds [7]
MBT [137]	-	-	-	-	-	-	-	52.3	-	85.0
CoDi [174]	-	-	74.4	-	-	-	-	-	78.9	-
ImageBind [69]*	50.0	-	-	-	-	-	-	27.8	-	-
BLIP-2 [109]*	-	-	-	9.2	18.3	-	36.7	-	-	-
InstructBLIP [34]*	-	-	-	22.1	41.8	-	38.3	-	-	-
Emu [172]**	-	-	-	24.1	39.8	-	-	-	-	-
Flamingo-9B [5]**	-	57.4	-	29.4	47.2	41.2	-	-	-	-
Flamingo-80B [5]	-	84.2	-	47.4	-	-	-	-	-	-
UIO-2 _L	68.5	37.1	44.0	39.6	48.2	51.0	37.5	37.8	45.7	86.1
UIO-2 _{XL}	71.4	41.6	47.1	39.3	50.4	52.0	45.6	44.2	45.7	88.0
UIO-2 _{XXL}	73.8	45.6	48.8	41.5	52.1	52.2	46.8	47.7	48.9	89.3

Inferences that stood the test of ALL evals.

Somewhat High Confidence Inferences:

1. GANs <<< Diffusion
 - Research idea: try UIO3: But with Diffusion .
2. Unimodal most likely wins when:
 - More "supervised" data + a few good "task-specific" tricks
 - DINO+SAM; LLAMA-7B
3. Unimodal most likely lose when:
 - Not enough data + No "additional" task-specific tricks
 - Mask-RCNN; GPV-½, CLIP

Shaky foundation Inferences:

1. Unimodal may or may not win when:
 - Not enough data, but a few "additional" tricks.
 - Wins -> AudioLDM; NLL-AngMF; VIMA
 - Losses -> AudioGen; DiffSound
 - Enough data, but no "additional" tricks.
 - Not the right experiments to analyze from this paper.

Let's Do Science, Not Just Alchemy:

1. Almost none of the inferences we eventually made, were "very" high confidence.
 - Yes, multi-modal models perform better when there's more data, and uni-modal models don't use tricks.
 - But Can we "pin-point" why?
 1. Representations Learned through -> Unification of modalities?
 2. More raw Data
 3. More raw Parameters
 - Yes, UIO-2 beats all other multi-modal models in Vision-Language tasks.
 - But Can we "pin-point" why?
2. Few "systematic", "controlled", "small-scale" experiments that "generalize" to large-scale, with "guarantees" would be a "much more" **scientific approach**.
3. Academics -> should have "MORE" incentive to do this, as it does not require heavy scale. But we seem to be stuck with the incentives from industry to just scale, without enough "resources" to go after it.

Limitations & Societal Implications

Limitations

- Long horizon Generation
 - Audio (limited to 4.08 s)
 - Video (perceiver Resampler bottleneck vs time)
- Image Quality (FID score is worse vs CoDi and Emu)
 - VQ-GAN intrinsic limitation
 - Its difficult to lower reconstruction loss with larger H*W input images.

Implications

- Pushes the efforts for unifying modalities into one model.
- 1 step towards a "unified", "open" interface for humans with general purpose AI.
- One model for n tasks is easier to scale vs n model for n tasks. Can lead to widespread availability at cheaper cost.
- Pessimistic side: enhances bad uses of deep fakes, now conditioned through any modality, and more faithful, albeit lower quality XD.

Summary of Strengths, Weaknesses, Relationships

Strengths

- Logistics of Unifying 4 modalities was very well handled.
 - Code, open data, experiments, tasks coverage, training regimes, qualitative samples generated.
- Tasks Evaluated were exhaustive, along with number of models compared against.
- Qualitative Samples shown were quite diverse, and exhaustive. Leads to a better intuition for weakness/strengths.
- Depth of explanation for stabilizing training with increasing modality was insightful !

Weakness

- Experiment analysis were more or less, a dump of numbers.
 1. Adding additional column for data seen per model for each table could help.
 2. Reasoning for under/over performance were not deep enough. (might be another paper altogether tho.)
- Extremely difficult to pin-point what works and what does not. In their defence, ablations are also quite infeasible at this scale. -> calls for controlled experiments in small-scale, and ways to discover llm laws that generalize to large-scale.

Thank You!

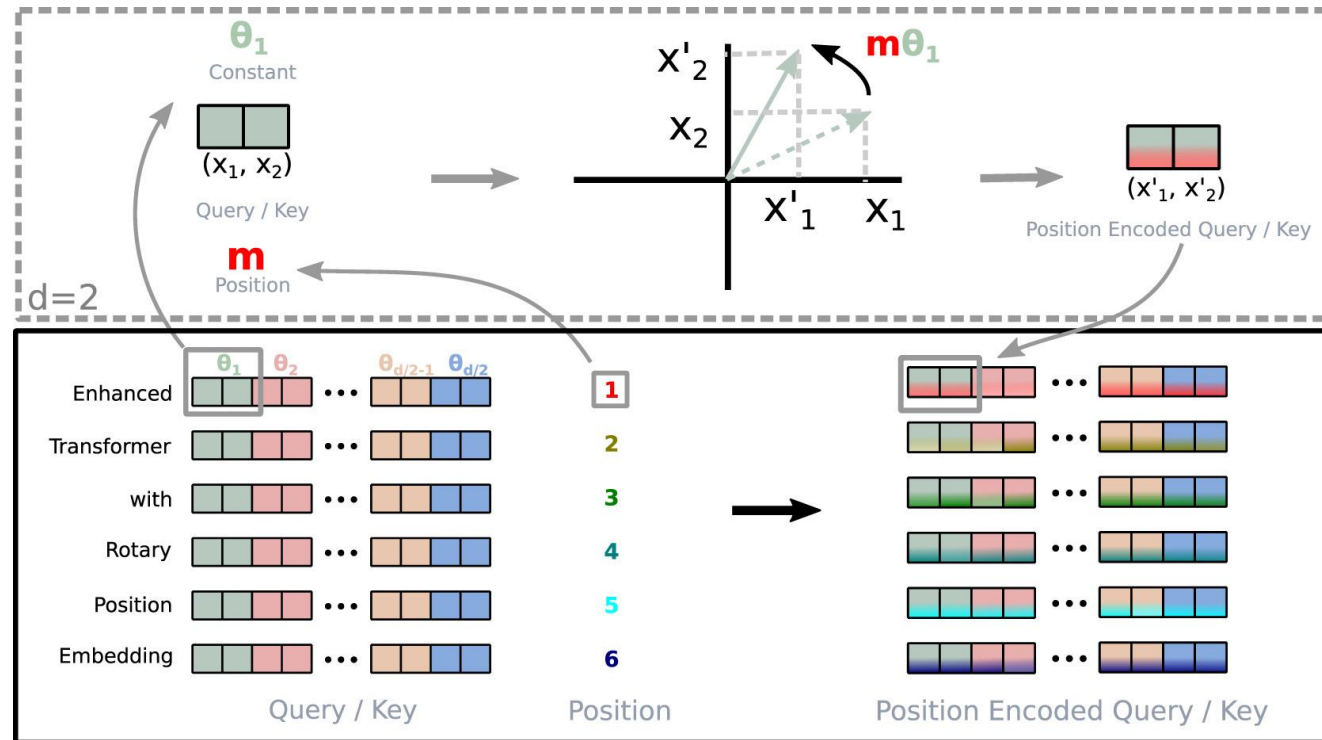
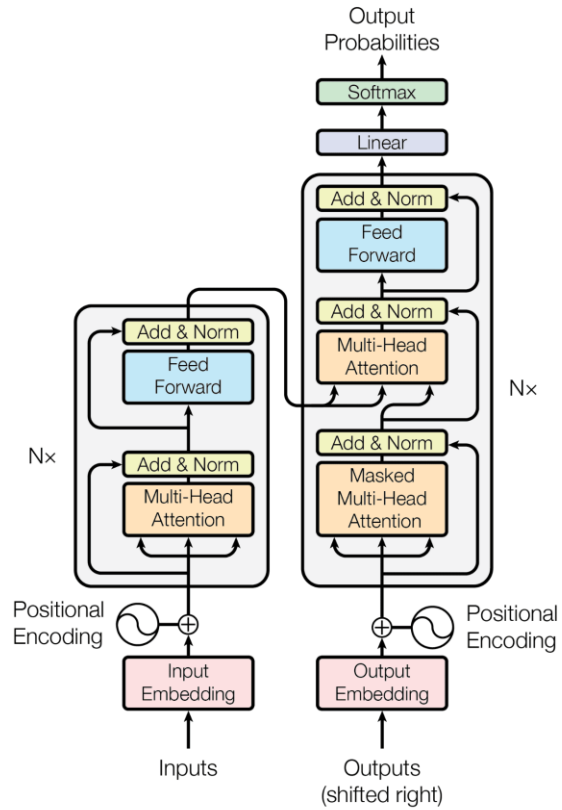
EXTRA

Architecture

Improvements

(1) RoPE

2D Rotary Positional Embeddings



RoPE

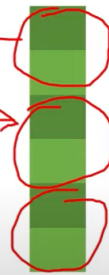
Implementation: Matrix Multiplication

$$f_{\{q,k\}}(\mathbf{x}_m, m) = \begin{pmatrix} \cos m\theta & -\sin m\theta \\ \sin m\theta & \cos m\theta \end{pmatrix} \begin{pmatrix} W_{\{q,k\}}^{(11)} & W_{\{q,k\}}^{(12)} \\ W_{\{q,k\}}^{(21)} & W_{\{q,k\}}^{(22)} \end{pmatrix} \begin{pmatrix} x_m^{(1)} \\ x_m^{(2)} \end{pmatrix}$$

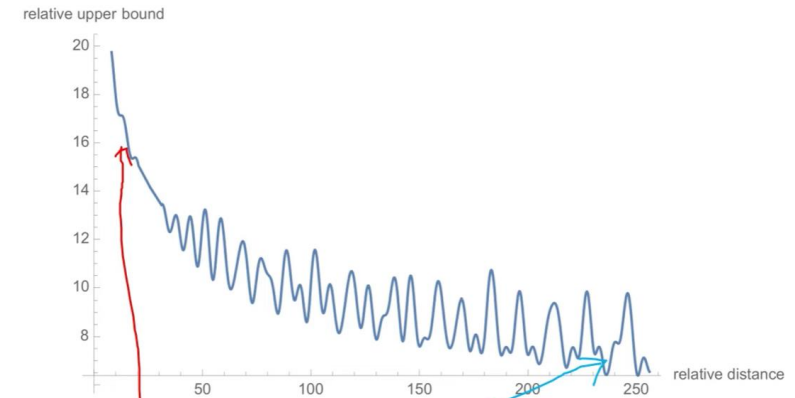
Q, K
not V

$$f_{\{q,k\}}(\mathbf{x}_m, m) = \mathbf{R}_{\Theta, m}^d \mathbf{W}_{\{q,k\}} \mathbf{x}_m$$

$$\mathbf{R}_{\Theta, m}^d = \begin{pmatrix} \cos m\theta_1 & -\sin m\theta_1 & 0 & 0 & \cdots & 0 & 0 \\ \sin m\theta_1 & \cos m\theta_1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cos m\theta_2 & -\sin m\theta_2 & \cdots & 0 & 0 \\ 0 & 0 & \sin m\theta_2 & \cos m\theta_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \cos m\theta_{d/2} & -\sin m\theta_{d/2} \\ 0 & 0 & 0 & 0 & \cdots & \sin m\theta_{d/2} & \cos m\theta_{d/2} \end{pmatrix}$$

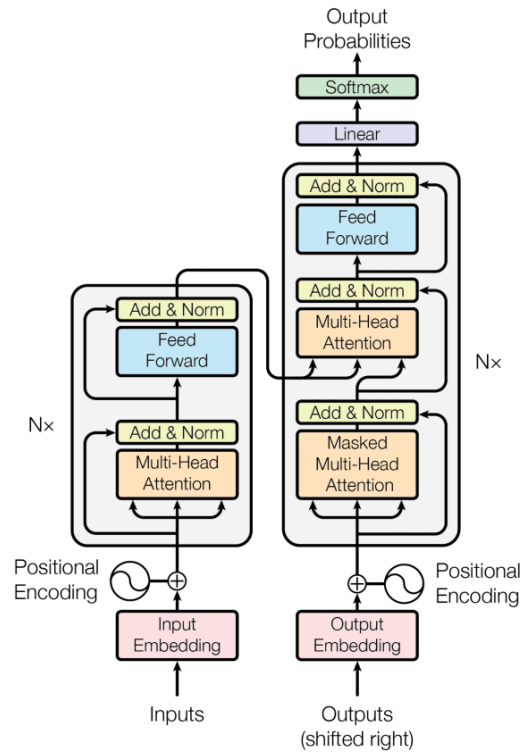


- Applied to Q,K and rotate each part of the vector by mtheta



Once upon a time, the pig chased the dog. And as the seasons turned and years drifted by, the pig and the dog lived their lives to the fullest, for they knew that the most magical journeys were the ones taken together. And as the final page of their adventure drew near, they knew that their story was one that would be whispered in the wind, celebrated by the stars, and cherished in the hearts of all who heard it. And they lived happily ever after.

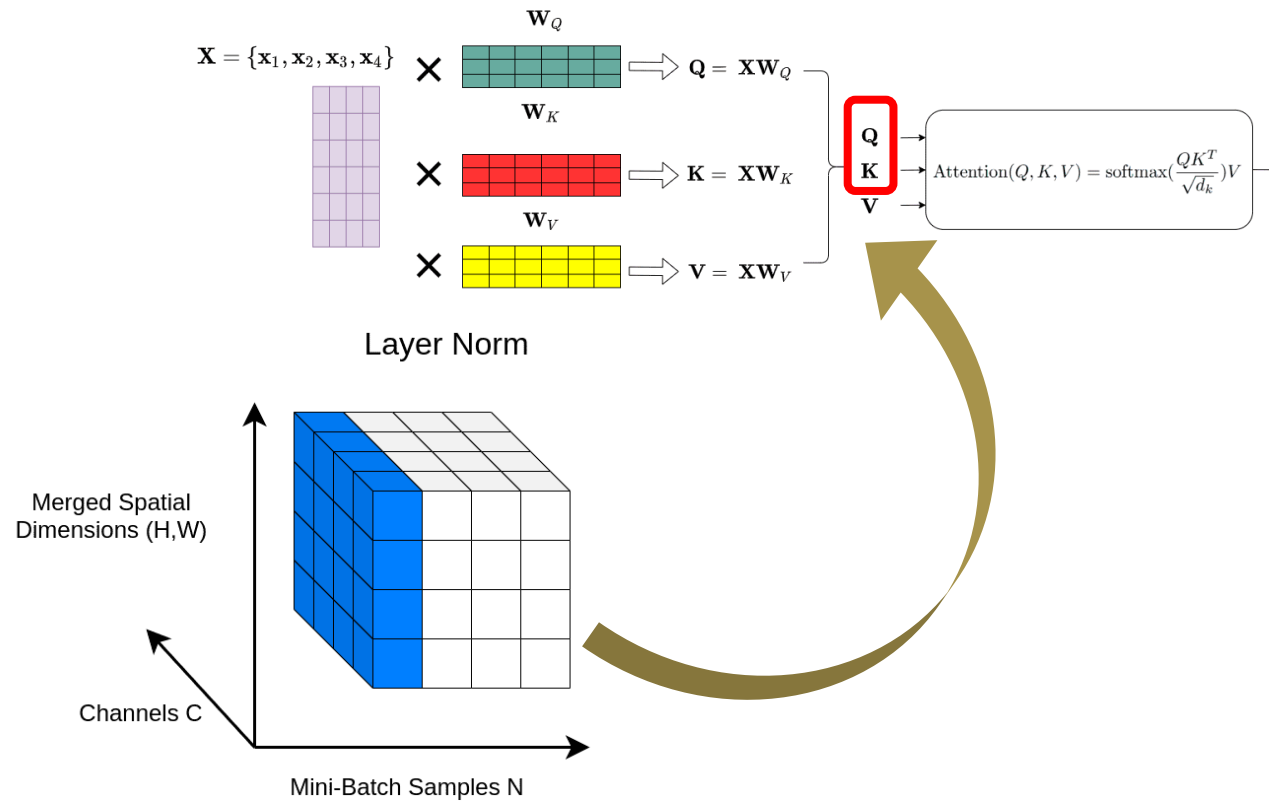
Architecture



Improvements

(1) 2D Rotary Positional Embeddings (RoPE)

(2) QK Normalization



LayerNorm

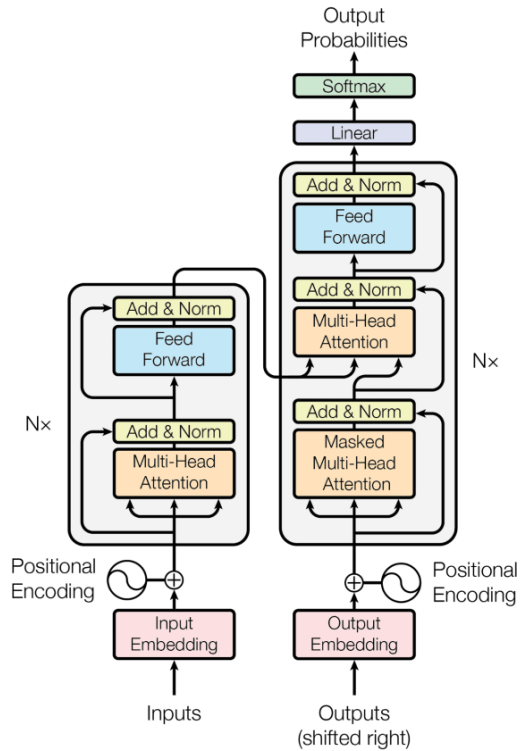
$$LN(x) = \gamma\left(\frac{x - \mu(x)}{\sigma(x)}\right) + \beta$$

$$\mu_n(x) = \frac{1}{CHW} \sum_{c=1}^C \sum_{h=1}^H \sum_{w=1}^W x_{nchw}$$

$$\sigma_n(x) = \sqrt{\frac{1}{CHW} \sum_{c=1}^C \sum_{h=1}^H \sum_{w=1}^W (x_{nchw} - \mu_n(x))^2}$$

- we can take the mean across the spatial dimension **and across all channels**

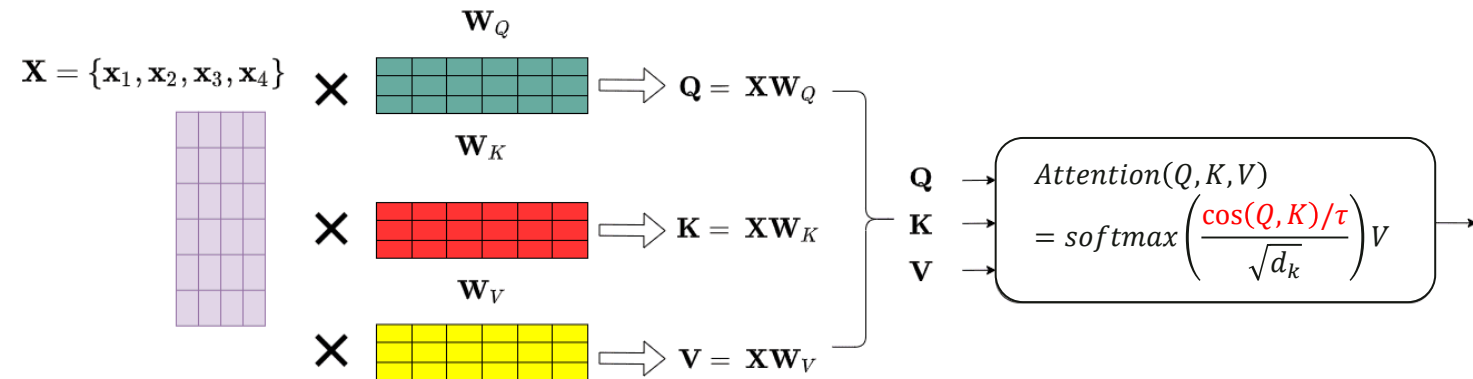
Architecture



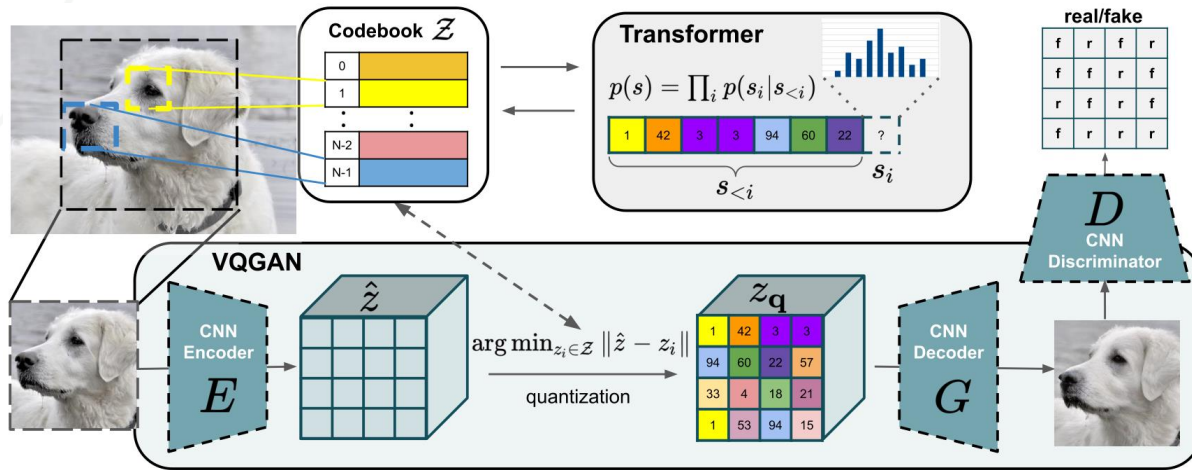
Improvements

- (1) 2D Rotary Positional Embeddings (RoPE)
- (2) QK Normalization
- (3) Scaled Cosine Attention

Perceiver Resampler



VQ-GAN

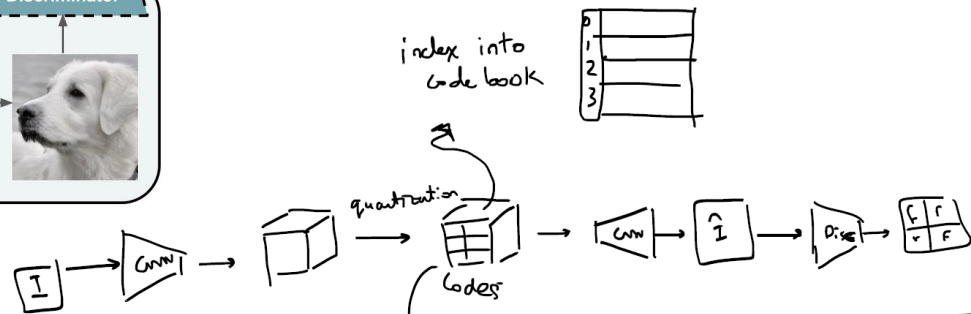


CNN + transformer

2-stage training

- ① GAN training to learn codebook
 - ② transformer to predict tokens from the codebook
- ↓
Sequence prediction task
autoregressively on codes

Freeze GAN

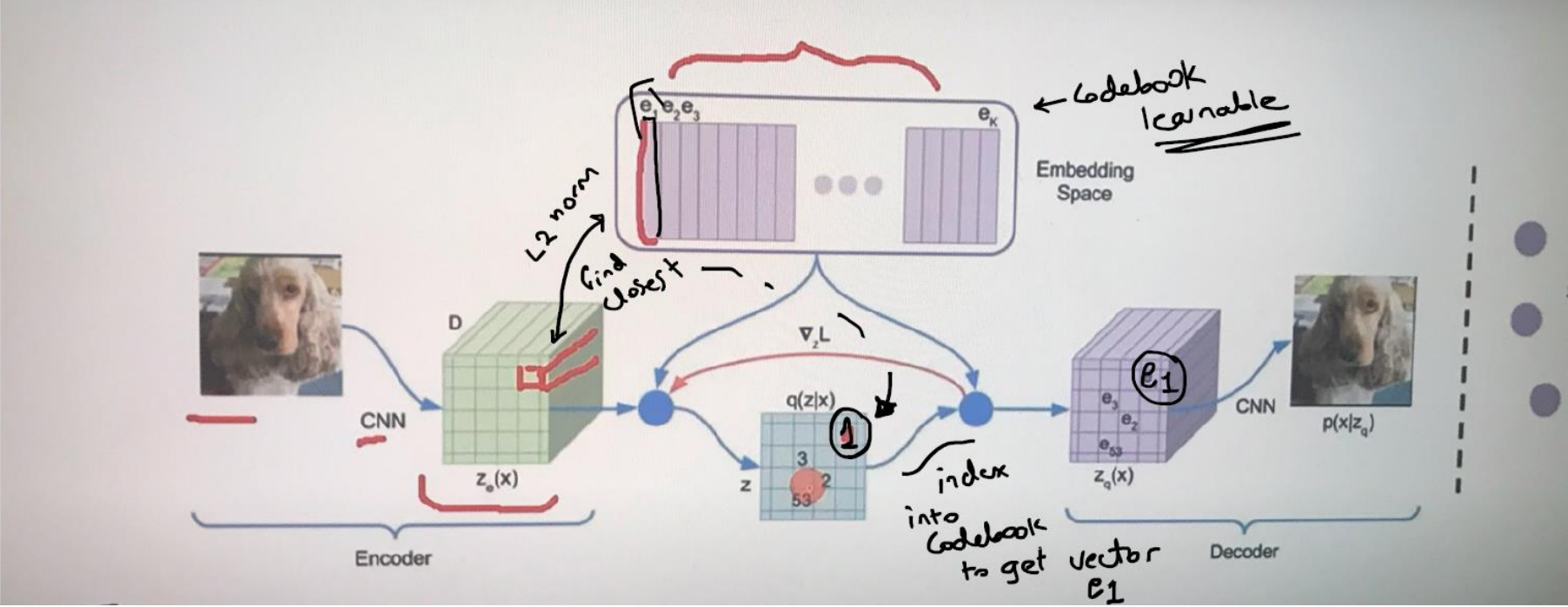


② take pred. from frozen encoder, then do $[5] [1] [?]$ pred. next code given prev.

To generate an image

use the transformer feed it $\langle s \rangle$ then it o/p/s $\langle 2 \rangle$
then feed it $\langle 2 \rangle$... autoregressively get long
seq. of tokens then reshape to 16×16 & feed into decoder !!

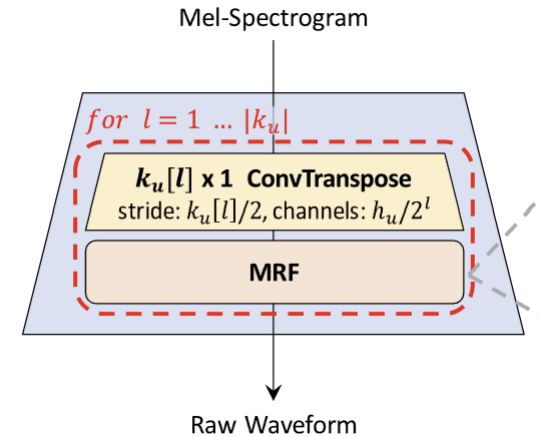
VQ-VAE



ViT VQ-GAN

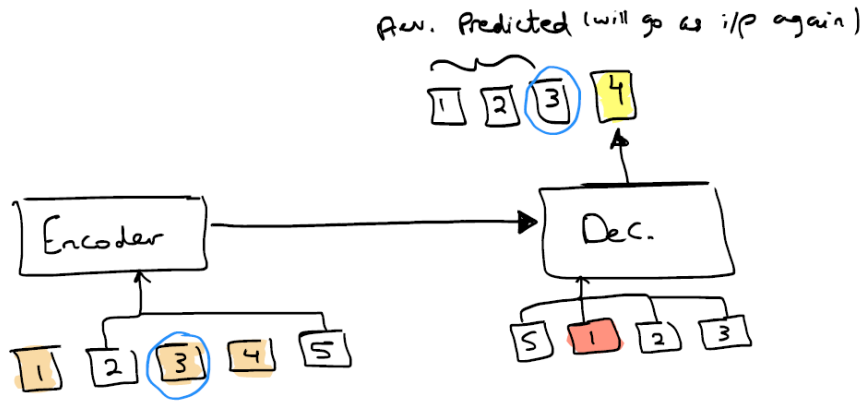
To generate audio, we use ViT-VQGAN [208] to convert the spectrograms into discrete tokens. Since the authors of [208] did not release the source code or any pre-trained models, we implement and train our own version of ViT-VQGAN with 8×8 patch size that encodes a 256×128 spectrogram into 512 tokens with a codebook size of 8196. The model is trained with the audio on AudioSet [54], ACAV100M [105], and YT-Temporal-1B [215] datasets. After getting the log-mel-scaled spectrograms, we use HiFi-GAN⁴ [98] vocoder to decode the spectrograms back to waveforms. We train the HiFi-GAN using the same parameters shown in Table 7. We trained the model on a mixture of AudioSet and LJSpeech [85] to cover natural sound and human voice.

HiFi-GAN



Pre-training Objectives

a

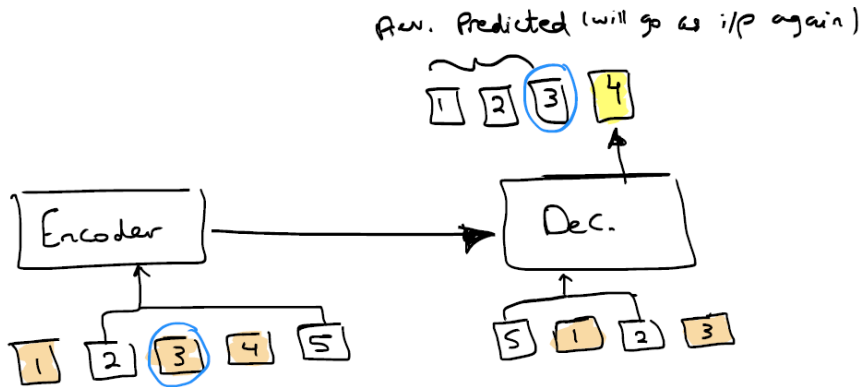


1 leaked even tho it was masked in the encoder as the decoder predicted it previously. (also 3)

- masked
- Decoder target → 4
- leaked as it was masked in encoder
- target

○ note that the decoder already predicted 1 & 3

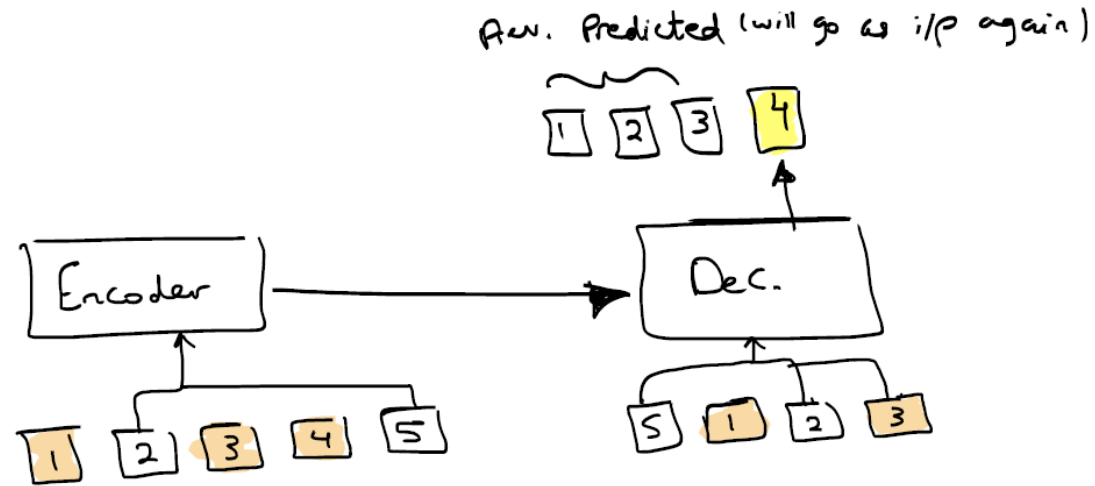
b



Whatever was masked in the encoder is masked in the decoder as well
 ↓ in generation ability when doing 50% masked vs 50% generation

Pre-training Objectives

©



masking the token unless when predicting it?

mostly removes data leakage while not interfering w/ causal prediction (generation)

Use current output [3] to predict next token [4]

Pre-Training Data

	Size	Rate	Text	Sparse	Dense	Image	Audio	ImageH	AudioH	Text	Sparse	Dense	Image	Audio
Video	181m	25.0	✓	-	-	✓	✓	✓	✓	✓	-	-	✓	✓
YT-Temporal [215]	146m	13.7	✓	-	-	✓	✓	✓	✓	✓	-	-	✓	✓
ACAV [105]	17m	3.98	✓	-	-	✓	-	✓	-	✓	-	-	✓	✓
HD-VILA [200]	7.1m	2.75	✓	-	-	✓	✓	✓	✓	✓	-	-	✓	✓
AudioSet [54]	1.7m	2.75	✓	-	-	✓	✓	✓	✓	✓	-	-	✓	✓
WebVid [13]	9.2m	1.23	✓	-	-	✓	✓	✓	✓	✓	-	-	✓	-
Ego4D [60]	0.7m	0.55	✓	-	-	✓	✓	✓	✓	✓	-	-	✓	✓
Interleaved Image/Text	157m	8.70	✓	-	-	✓	-	✓	-	✓	-	-	✓	-
OBELICS [104]	131m	8.00	✓	-	-	✓	-	✓	-	✓	-	-	✓	-
CC12M Interleaved	11m	0.35	✓	-	-	✓	-	✓	-	✓	-	-	-	-
CC3M Interleaved	3.0m	0.21	✓	-	-	✓	-	✓	-	✓	-	-	-	-
RedCaps Interleaved	12m	0.14	✓	-	-	✓	-	✓	-	✓	-	-	-	-
Multi-View	3.4m	0.67	✓	-	-	✓	-	✓	-	-	✓	-	✓	-
CroCo Habitat [157, 194]	2.6m	0.33	✓	-	-	✓	-	✓	-	-	-	-	✓	-
Objaverse [40]	0.8m	0.33	✓	-	-	✓	-	✓	-	-	✓	-	✓	-

Pre-Training Data

	Size	Rate	Text	Sparse	Dense	Image	Audio	ImageH	AudioH	Text	Sparse	Dense	Image	Audio
Agent Trajectories	1.3m	0.33	✓	-	-	✓	-	✓	-	✓	-	-	✓	-
ProcTHOR [39]	0.7m	0.17	✓	-	-	✓	-	✓	-	✓	-	-	✓	-
Habitat [157]	0.6m	0.17	✓	-	-	✓	-	✓	-	✓	-	-	✓	-
Synthetic	504m	1.00	✓	✓	-	✓	-	-	-	-	✓	✓	-	-
Segment Anything [94]	1.1m	0.50	✓	✓	-	✓	-	-	-	-	-	✓	-	-
Laion Aesthetics Patches	491m	0.45	✓	-	-	✓	-	-	-	-	✓	-	-	-
RedCaps Patches	12m	0.05	✓	-	-	✓	-	-	-	-	✓	-	-	-
All	8.5b	100	✓	✓	-	✓	✓	✓	✓	✓	✓	✓	✓	✓

Instruction Tuning Data

- **220 tasks** from over **120 datasets**

	Size	Rate	Datasets	Text	Sparse	Dense	Image	Audio	ImageH	AudioH	Text	Sparse	Dense	Image	Audio
Image Generation	506m	17.6	21	✓	✓	✓	✓	-	✓	✓	✓	-	-	✓	-
Image from Text	497m	10.6	5	✓	-	-	-	-	-	-	-	-	-	✓	-
Controllable Image Editing	3.0m	2.92	4	✓	-	✓	✓	-	✓	-	-	-	-	✓	-
Image Editing	1.1m	1.66	3	✓	-	-	✓	-	-	-	-	-	-	✓	-
Next Frame Generation	24k	0.96	2	✓	✓	-	-	-	✓	✓	-	-	-	✓	-
Image Inpainting	1.0m	0.79	3	✓	✓	-	✓	-	-	-	-	-	-	✓	-
View Synthesis	4.2m	0.60	4	✓	-	-	✓	-	✓	-	✓	-	-	✓	-
Audio Generation	164m	7.50	9	✓	-	-	✓	✓	✓	✓	-	-	-	-	✓
Audio from Text	19m	5.62	8	✓	-	-	-	-	-	✓	-	-	-	-	✓
Audio from Video	145m	1.88	1	✓	-	-	✓	✓	✓	✓	-	-	-	-	✓
Image Understanding	53m	17.8	73	✓	✓	-	✓	-	✓	-	✓	-	-	-	-
VQA	5.8m	6.23	31	✓	-	-	✓	-	-	-	✓	-	-	-	-
Image Captioning	32m	4.25	14	✓	-	-	✓	-	-	-	✓	-	-	-	-
Region Classification	6.1m	2.41	4	✓	✓	-	✓	-	-	-	✓	-	-	-	-
Image Tagging	3.8m	2.38	8	✓	-	-	✓	-	-	-	✓	-	-	-	-
Relationship Prediction	0.8m	1.41	6	✓	✓	-	✓	-	-	-	✓	-	-	-	-
Region Captioning	3.5m	0.60	1	✓	✓	-	✓	-	-	-	✓	-	-	-	-
Image Instruction Following	0.4m	0.37	6	✓	-	-	✓	-	-	-	✓	-	-	-	-
Image Pair QA	0.1m	0.17	3	✓	-	-	✓	-	✓	-	✓	-	-	-	-

Instruction Tuning Data

	Size	Rate	Datasets	Text	Sparse	Dense	Image	Audio	ImageH	AudioH	Text	Sparse	Dense	Image	Audio
Image Sparse Labelling	13m	7.25	26	✓	✓	-	✓	-	✓	-	-	✓	-	✓	-
Object Detection	5.3m	3.08	9	✓	-	-	✓	-	-	-	-	✓	-	-	-
Object Localization	6.0m	1.31	3	✓	-	-	✓	-	-	-	-	✓	-	-	-
Referring Expression	0.2m	1.08	7	✓	-	-	✓	-	-	-	-	✓	-	-	-
3D	1.0m	1.00	2	✓	-	-	✓	-	✓	-	-	✓	-	✓	-
Text Detection	37k	0.41	3	✓	-	-	✓	-	-	-	-	✓	-	-	-
Keypoint Detection	0.3m	0.38	2	✓	✓	-	✓	-	-	-	-	✓	-	-	-
Image Dense Labelling	6.9m	4.06	19	✓	✓	-	✓	-	✓	-	-	-	✓	-	-
Semantic Segmentation	2.4m	1.23	4	✓	-	-	✓	-	-	-	-	-	✓	-	-
Localized Segmentation	3.2m	1.17	3	✓	✓	-	✓	-	-	-	-	-	✓	-	-
Surface Normal Estimation	1.1m	1.03	6	✓	-	-	✓	-	-	-	-	-	✓	-	-
Referring Expression Segmentation	0.1m	0.47	3	✓	-	-	✓	-	-	-	-	-	✓	-	-
Depth Estimation	47k	0.11	1	✓	-	-	✓	-	-	-	-	-	✓	-	-
Optical Flow	24k	0.06	2	✓	-	-	✓	-	✓	-	-	-	✓	-	-
Video Understanding	13m	10.6	24	✓	-	-	✓	✓	✓	✓	✓	-	-	-	-
Video Captioning	9.1m	3.75	3	✓	-	-	✓	-	✓	✓	✓	-	-	-	-
Video Tagging	1.1m	3.75	6	✓	-	-	✓	-	✓	✓	✓	-	-	-	-
Video Question Answering	2.5m	2.84	9	✓	-	-	✓	✓	✓	✓	✓	-	-	-	-
Video Instruction Following	0.2m	0.21	6	✓	-	-	✓	-	✓	✓	✓	-	-	-	-

Instruction Tuning Data

	Size	Rate	Datasets	Text	Sparse	Dense	Image	Audio	ImageH	AudioH	Text	Sparse	Dense	Image	Audio
Video Sparse Labelling	0.4m	3.42	5	✓	✓	-	✓	-	✓	✓	-	✓	-	-	-
Video Tracking	0.2m	2.50	3	✓	✓	-	✓	-	✓	✓	-	✓	-	-	-
Video Action Localization	0.2m	0.61	1	✓	-	-	✓	-	✓	✓	-	✓	-	-	-
Video Sound Localization	2.5k	0.31	1	✓	-	-	✓	-	✓	✓	-	✓	-	-	-
Audio Understanding	2.2m	2.50	10	✓	-	-	✓	✓	✓	-	✓	-	-	-	-
Audio Tagging	2.1m	1.25	5	✓	-	-	✓	✓	✓	-	✓	-	-	-	-
Audio Captioning	75k	1.25	5	✓	-	-	-	✓	-	-	✓	-	-	-	-
Natural Language	11m	25.0	17	✓	-	-	-	-	-	-	✓	-	-	-	-
Text Instruction Following	11m	12.5	10	✓	-	-	-	-	-	-	✓	-	-	-	-
Language Modeling	-	12.5	7	✓	-	-	-	-	-	-	✓	-	-	-	-
Embodied AI	7.2m	4.33	23	✓	-	-	✓	-	✓	-	✓	✓	-	✓	-
Action Prediction	4.3m	3.37	12	✓	-	-	✓	-	✓	-	✓	-	-	-	-
Next Frame/State Prediction	1.3m	0.33	2	✓	-	-	✓	-	✓	-	✓	-	-	✓	-
Goal Generation	0.7m	0.33	3	✓	-	-	✓	-	✓	-	-	-	-	✓	-
Embodied QA	1.0m	0.30	6	✓	-	-	✓	-	✓	-	✓	✓	-	-	-
All Tasks	775m	100	227	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

- 60% data with prompts
- Catastrophic forgetting → 30% of the data is carried over from pre-training