

Kosmos-1/2 & UnifiedIO v1/2

Jay Javeri, Jingyang Ke, Neel Shah

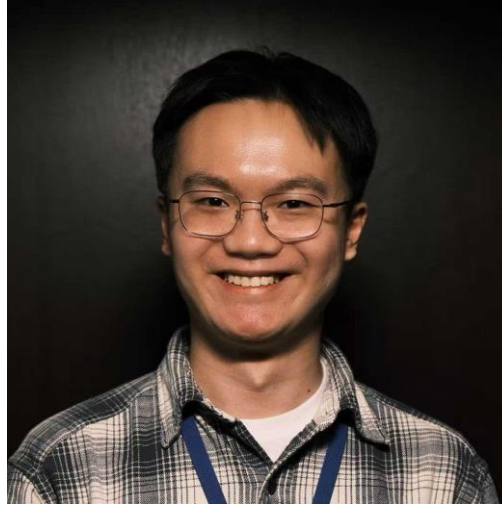
Introductions



Jay Javeri

1st Year - MSCS ML

Areas of Interest :
Reasoning in Language
models, Multi-modality
and Efficient Machine
Learning



Jingyang Ke

3rd year - ML PhD

Areas of interest:
Reinforcement
Learning, NeuroAI,
Multimodal LLMs



Neel Shah

3rd year - Robotics PhD

Areas of interest:
sensor design, controls,
additive Manufacturing

Outline

AI2 Allen Institute for AI

UnifiedIO-1

Unified “any-to-any” architecture

2022



Kosmos-1

OCR-free vision from scratch

Early-2023



Kosmos-2

Grounded text/vision

Mid-2023

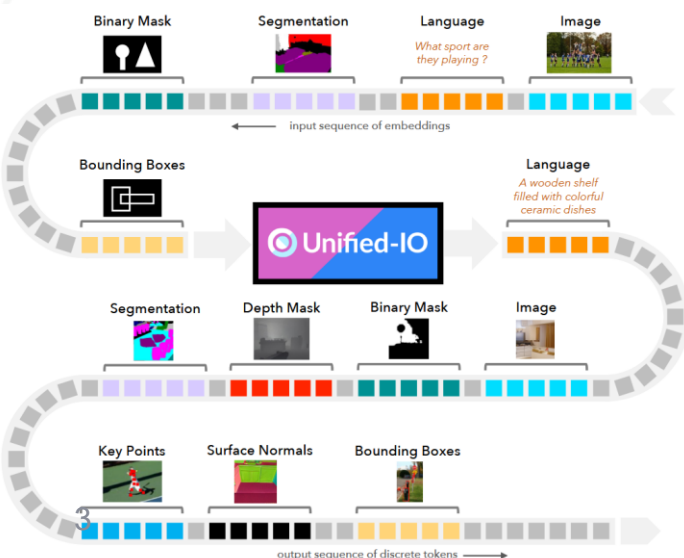
Kosmos-G

AI2 Allen Institute for AI

UnifiedIO-2

Expanded task repertoire + prompting

late-2023



Text
Vision
Documents

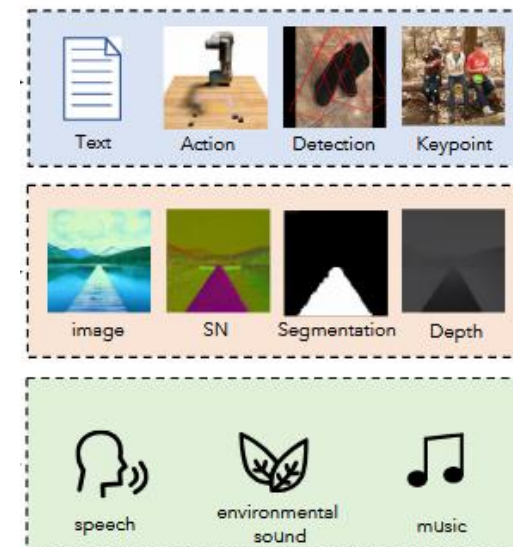
→

Text

Text
Vision
B. Boxes

→

Text
B.
Boxes



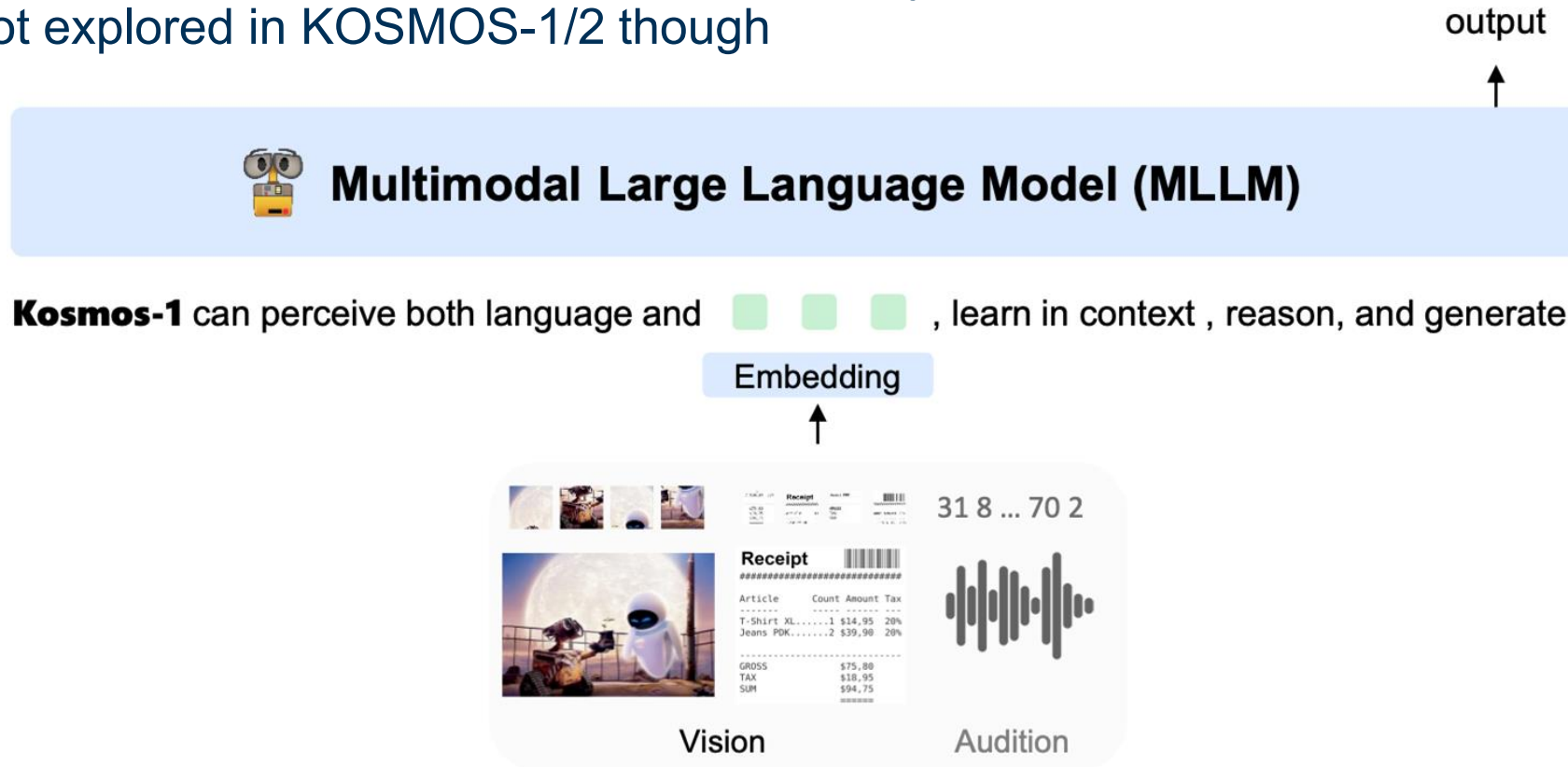
Language Is Not All You Need: Aligning Perception with Language Models (KOSMOS-1)

(Huang et al., NeurIPS 2023)

- Multimodal Large Language Model framework
- Pretrained from scratch

Architecture

- KOSMOS-1 has a standard Transformer-based causal language model architecture
 - Model size: 1.6B parameters
- Can be extended to other modalities beyond vision
 - Not explored in KOSMOS-1/2 though



MAGNETO

- Extra LayerNorm to each sublayer
- Better training stability and superior performance across modalities.

Models			Previous	This work
Vision	Encoder	ViT/BEiT	Pre-LN	Sub-LN
Language	Encoder	BERT	Post-LN	
	Decoder	GPT	Pre-LN	
	Encoder-Decoder	NMT/BART	Post-LN	
Speech	Encoder	T-T	Pre-LN	Sub-LN
Multimodal	Encoder	BEiT-3	Pre-LN	

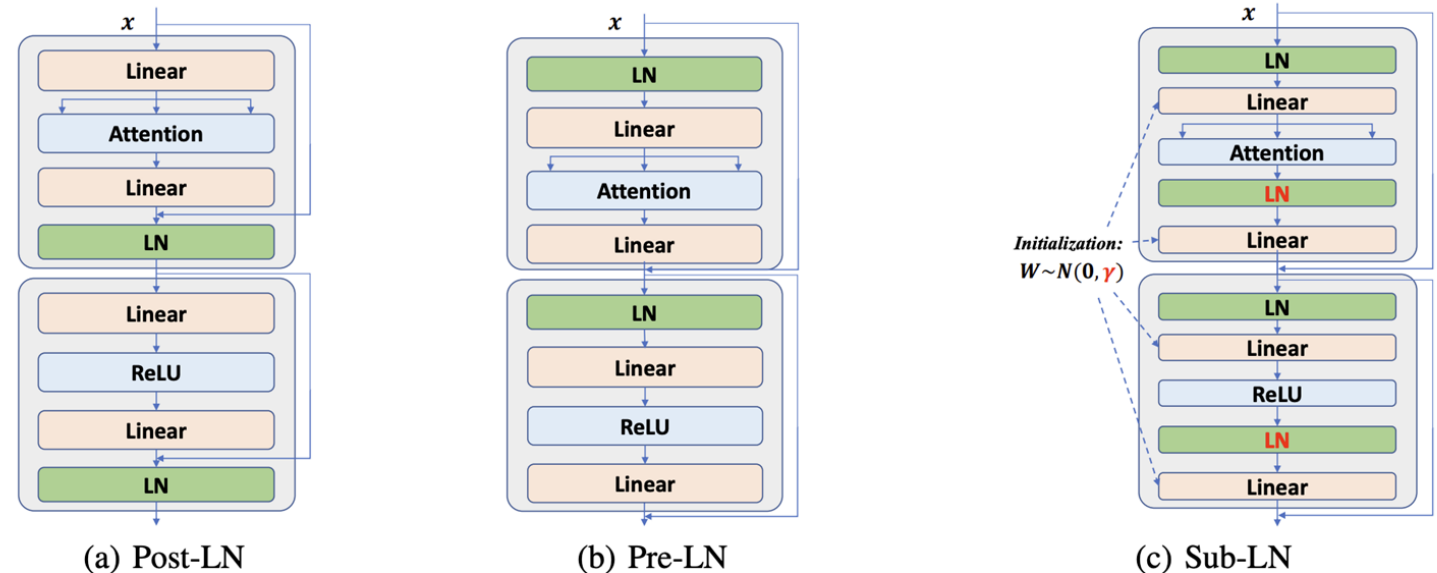


Figure 1: **Top:** the architectures of SOTA models across language, vision, speech, and multimodal. **Bottom:** the proposed Foundation Transformer uses Sub-LN and theoretically derived initialization.

Extrapolatable Position Embedding (XPOS)

- Optimizes attention resolution so that the position information can be captured more precisely
 - usual pairwise rotation (RoPE) + per-dimension exponential scaling for Q and K
- Block-wise causal attention in inference
- Generalize to different lengths better

Training

Pretrain with a mix of following datasets

- Text Corpora
 - Several massive datasets for training LLMs
- Image-Caption Pairs
 - English LAION-2B, LAION-400M, and COYO-700M, Conceptual Captions
 - Image-caption datasets from internet web pages
- Interleaved Image-Text Data
 - 71M web pages from the Common Crawl snapshot
 - Extract the text and images from the HTML of each selected web page

Language-only instruction tuning

- Train the model with the instruction data in the format of (instructions, inputs, and outputs)

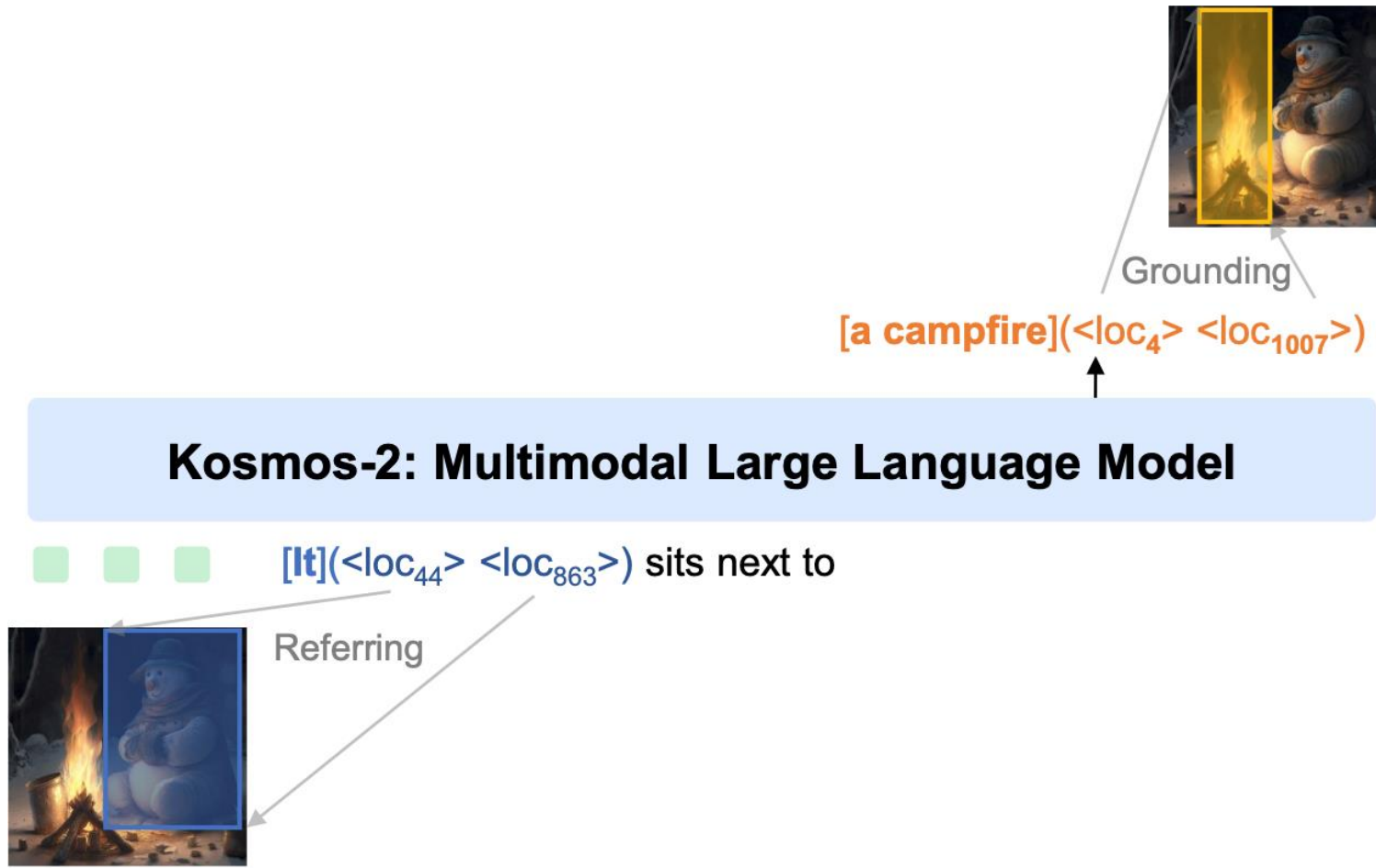
Grounding Multimodal Large Language Models to the World (KOSMOS-2)

(Peng et al., ICLR 2024)

- Add grounding to a general-purpose MLLM

KOSMOS-2

- Add grounding (linking text \leftrightarrow image regions) to a general-purpose MLLM
- Same architecture as KOSMOS-1, same model size (1.6B)



Why Grounding?

- MLLMs “see” & “talk,” but often can’t point (refer to concrete regions)
- Many tasks need region-aware reasoning (phrase grounding, referring, grounded VQA/captioning)
- What we want: unified, scalable grounding without custom detectors

Input Prompt



the left eye of emoji



Question: How many cows are here? Answer:



Question: What does the sign say? Answer:

Completion



(1)

two cows present in the image.



(2)

The sign says "Welcome to Carnaby Street."



(3)

Input Prompt



Question: Why is this animal unusual? Answer:



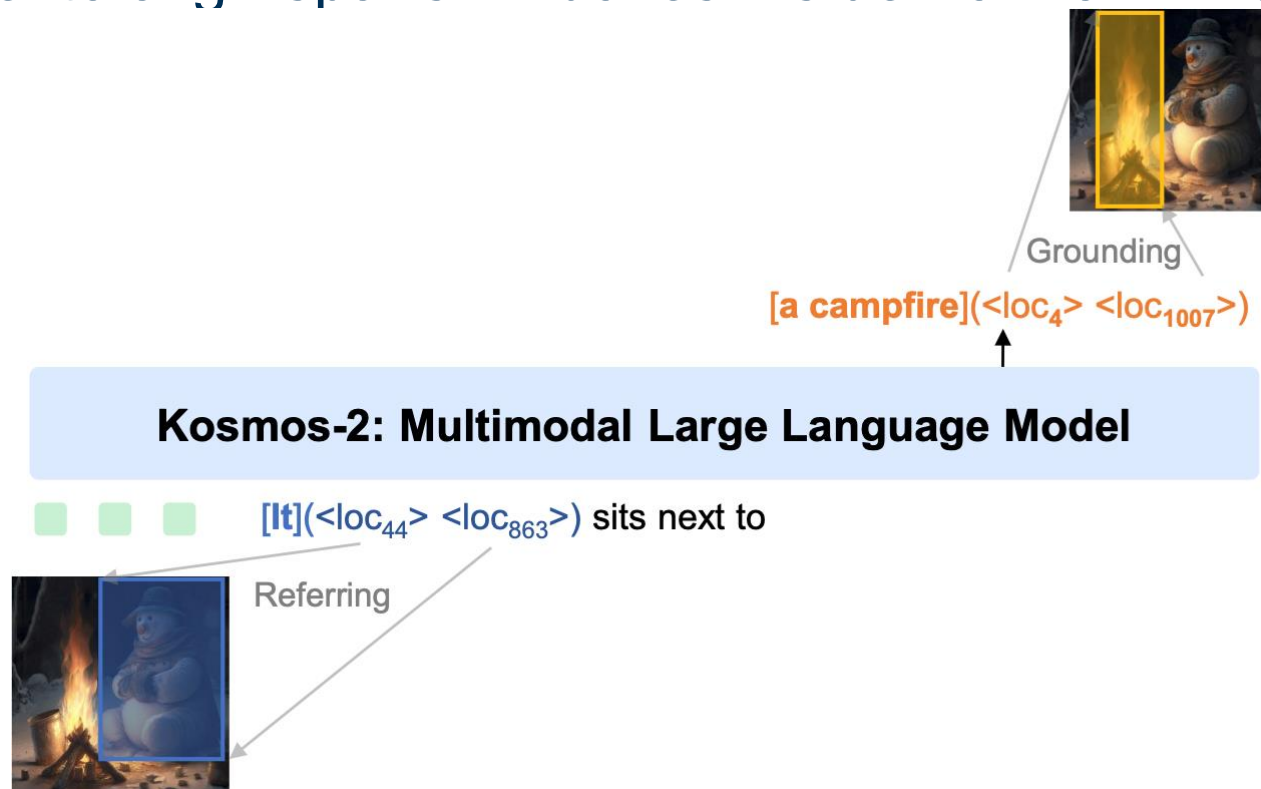
Question: What is it? Answer:



Question: What is the biggest difference between bottle-1 and bottle-2? Answer:

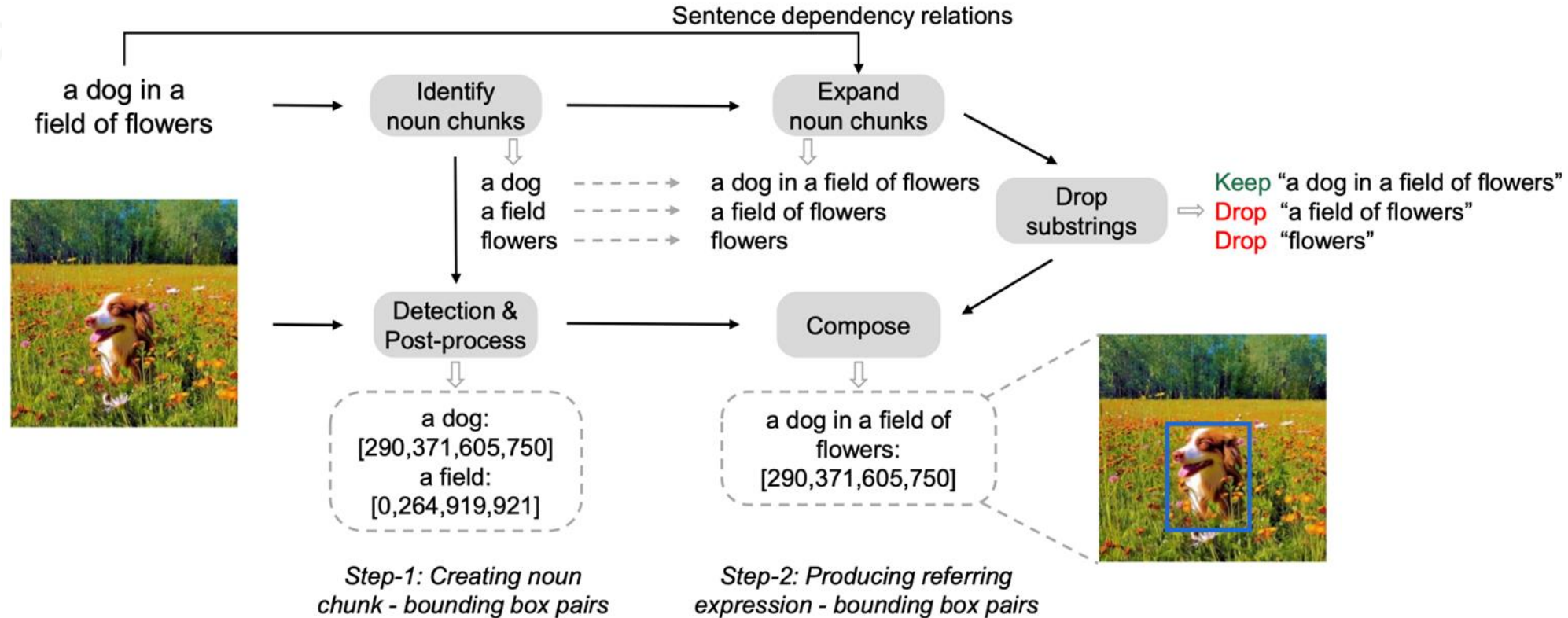
Grounded Input Representations

- Represent references as Markdown-style links from text spans → location tokens
- Discretize image into $P \times P$ bins; each bin has a special token $\langle loc_i \rangle$
- A bounding box = $\langle box \rangle \langle loc_1 \rangle \langle loc_2 \rangle \langle /box \rangle$ (top-left + bottom-right)
- Train the model to align spans \leftrightarrow boxes inside normal LM decoding



Web-Scale Grounded Image-Text Pairs (GRIT)

- Built on image-text pairs from subsets of LAION-2B & COYO-700M
- A two-step pipeline to extract and link text spans in the caption to their corresponding image regions



Training

Pretrain with a mix of below datasets

- Same as KOSMS-1
 - Text Corpora
 - Image-Caption Pairs
 - Interleaved Image-Text Data
- New: Grounded pairs (GRIT)

Fine-tuning

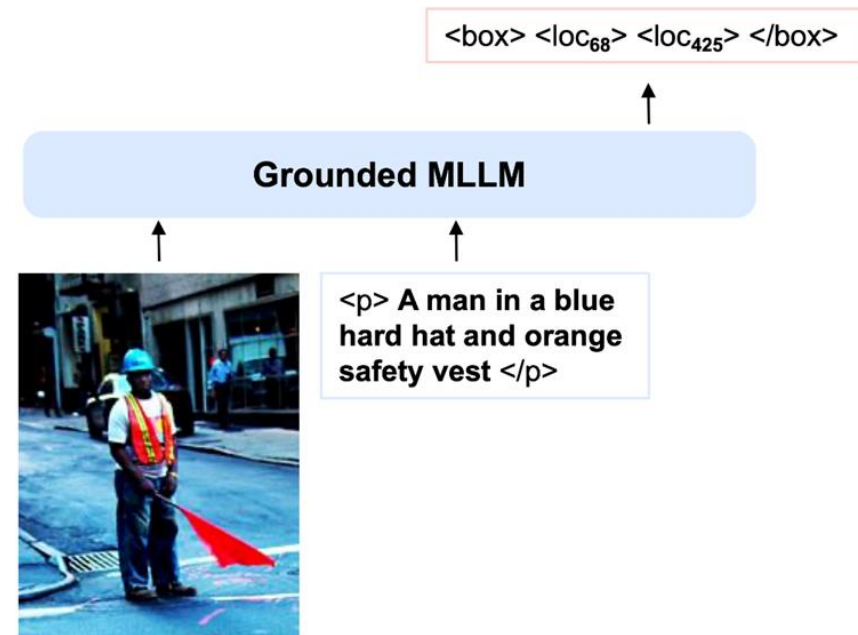
- Combine vision-language instruction dataset and language-only instruction datasets
 - LLaVA-Instruct
 - Unnatural Instructions & FLANv2
- Additional grounded instruction data by utilizing the pairs of bounding boxes and expressions in GRIT
 - Prompt the model to generate the corresponding location tokens or expressions of the bounding boxes

Results: Multimodal Grounding

- Phrase Grounding
 - LLaVA-Instruct
 - Unnatural Instructions & FLANv2
- Referring Expression Comprehension
 - prompt the model to generate the corresponding location tokens or expressions of the bounding boxes



(1) Phrase grounding



(2) Referring expression comprehension

Results: Multimodal Grounding

Model	Zero-shot	Val Split			Test Split		
		R@1	R@5	R@10	R@1	R@5	R@10
VisualBert [LYY ⁺ 19]	✗	70.4	84.5	86.3	71.3	85.0	86.5
MDETR [KSL ⁺ 21]	✗	83.6	93.4	95.1	84.3	93.9	95.8
GLIP [LZZ ⁺ 22]	✗	86.7	96.4	97.9	87.1	96.9	98.1
FIBER [DKG ⁺ 22]	✗	87.1	96.1	97.4	87.4	96.4	97.6
GRILL [JMC ⁺ 23]	✓	-	-	-	18.9	53.4	70.3
KOSMOS-2	✓	77.8	79.2	79.3	78.7	80.1	80.1

Table 2: Phrase grounding results on Flickr30k Entities. We report the R@1, R@5, and R@10 metrics, where R@1/5/10 means calculating the recall using the top 1/5/10 generated bounding boxes.

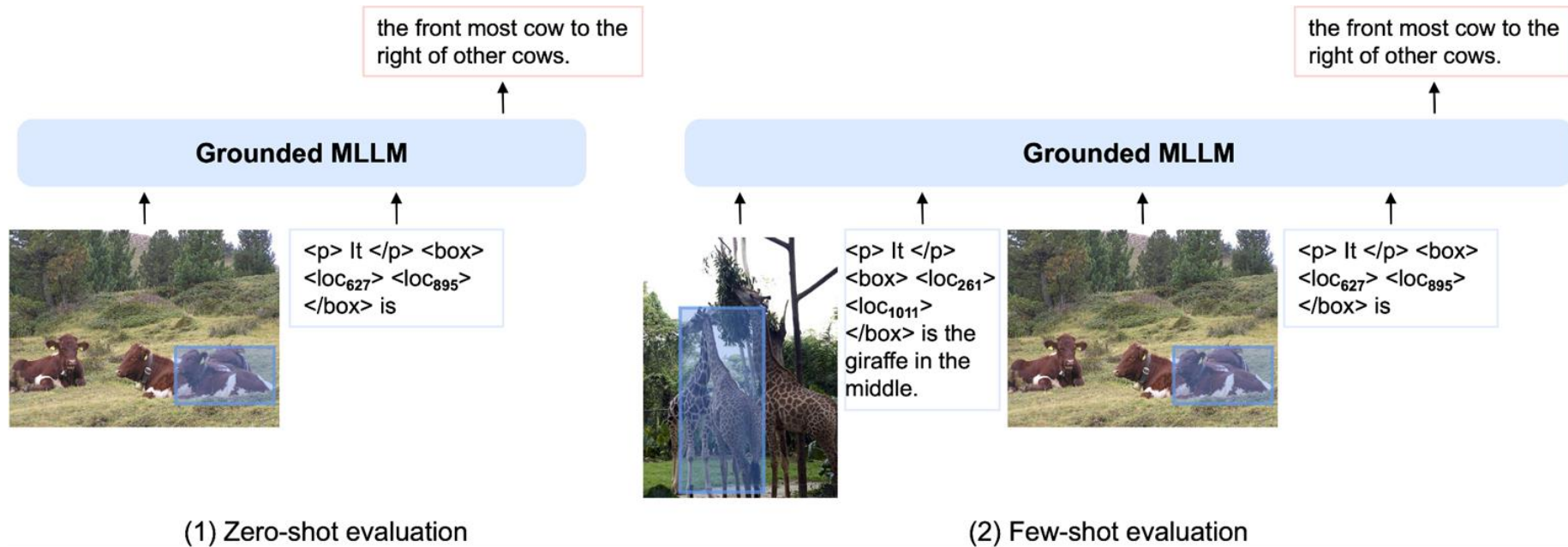
- Large improvement in zero-shot performance
- Still a gap between KOSMOS-2 and non-zero-shot methods
 - Especially in referring expression comprehension

Model	Zero-shot	RefCOCO			RefCOCO+			RefCOCOg	
		val	testA	testB	val	testA	testB	val	test
UNITER [CLY ⁺ 19]	✗	81.41	87.04	74.17	75.90	81.45	66.70	74.86	75.77
MDETR [KSL ⁺ 21]	✗	87.51	90.40	82.67	81.13	85.52	72.96	83.35	83.31
OFA [WYM ⁺ 22]	✗	90.05	92.93	85.26	84.49	90.10	77.77	84.54	85.20
FIBER [DKG ⁺ 22]	✗	90.68	92.59	87.26	85.74	90.13	79.38	87.11	87.32
VisionLLM [WCC ⁺ 23]	✗	86.7	-	-	-	-	-	-	-
GRILL [JMC ⁺ 23]	✓	-	-	-	-	-	-	-	47.5
KOSMOS-2	✓	52.32	57.42	47.26	45.48	50.73	42.24	60.57	61.65

¹⁶ Table 3: Referring expression comprehension results on RefCOCO, RefCOCO+ and RefCOCOg. We report the accuracy metric for all methods.

Results: Multimodal Referring

- Understand the region referred via input bounding boxes



Results: Multimodal Referring

- Impressive zero-shot performance
 - Even outperform finetuned SLR on CIDEr

Model	Setting	RefCOCOg	
		Meteor	CIDEr
SLR[YTBB17]	Finetuning	15.4	59.2
SLR+Rerank[YTBB17]	Finetuning	15.9	66.2
KOSMOS-2	Zero-shot	12.2	60.3
	Few-shot ($k = 2$)	13.8	62.2
	Few-shot ($k = 4$)	14.1	62.3

Table 4: Results of referring expression generation on RefCOCOg.

Results: General VL Tasks

- Flickr30k: Image captioning
- VQAv2: visual question-answering
 - Why KOSMOS-2 performance drop a little bit?

Model	Flickr30k	VQAv2
	CIDEr	VQA acc.
FewVLM [JCS ⁺ 22]	31.0	-
METALM [HSD ⁺ 22]	43.4	41.1
Flamingo-3B [ADL ⁺ 22]	60.6	49.2
Flamingo-9B [ADL ⁺ 22]	61.5	51.8
KOSMOS-1	65.2	46.7
KOSMOS-2	66.7	45.6

Table 5: Zero-shot image captioning results on Flickr30k test set and zero-shot visual question answering results on VQAv2 test-dev set. We report results of KOSMOS-2 and KOSMOS-1 without instruction tuning.

Results: Language-only Tasks

- Overall similar performance as LLM & KOSMOS-1
- BoolQ (T/F QA)
 - KOSMOS-2 achieves better results compared to LLM & KOSMOS-1
- CB (CommitmentBank)
 - Understand speaker commitment to the truth of a clause
 - KOSMOS-1 shows improvement but KOSMOS-2 has much worse performance

Model	Story Cloze	Hella Swag	Winograd	Winogrande	PIQA	BoolQ	CB	COPA
LLM	72.9	50.4	71.6	56.7	73.2	56.4	39.3	68.0
KOSMOS-1	72.1	50.0	69.8	54.8	72.9	56.4	44.6	63.0
KOSMOS-2	72.0	49.4	69.1	55.6	72.9	62.0	30.4	67.0

Table 6: Zero-shot performance comparisons of language tasks between KOSMOS-2, KOSMOS-1 and LLM. LLM uses the same text data and training setup to reimplement a language model as KOSMOS-1. We report results of KOSMOS-2 and KOSMOS-1 without instruction tuning. Results of

Discussion

Strengths:

- Elegant token-level grounding inside a standard VLM decoder
- Strong zero-shot performance on grounding tasks
- Effective GrIT pipeline to preprocess massive grounding data for training

Weaknesses:

- Limited novelty in architecture
- Model size (1.6B) is small, limiting its generalization ability
- Slight drop of KOSMOS-2 on VQA vs KOSMOS-1 suggests trade-offs in training mix

Unified-IO: A Unified Model for Vision, Language, and Multi-Modal Tasks (Unified IO 1)

(Lu et al., 2022 – Allen Institute for AI)

- First attempt at a single Seq2Seq model for a very wide set of AI tasks
- Trained jointly on 90+ datasets, 95 tasks
- Handles vision, language, and vision+language tasks

Motivation

Why Unified-IO?

Traditional CV models: task-specific heads (e.g. Mask R-CNN, VQA models)

NLP: success of Seq2Seq token-based models (T5, GPT-3)

Challenge: Vision outputs are very different (boxes, masks, depth maps, images)

Goal: Homogenize everything into tokens → single transformer can learn all tasks

Architecture

Base: T5-style Transformer encoder–decoder

All inputs/outputs → sequences of tokens from a shared vocabulary

Text: SentencePiece tokens

Dense outputs (images, masks, depth, normals): encoded into tokens via VQ-VAE

Sparse outputs (boxes, keypoints): encoded as coordinate tokens

Vocabulary: ~50k tokens (32k text, 16k image, 1k location)

UnifiedIO-1

What are the risks of forcing everything into discrete tokens?

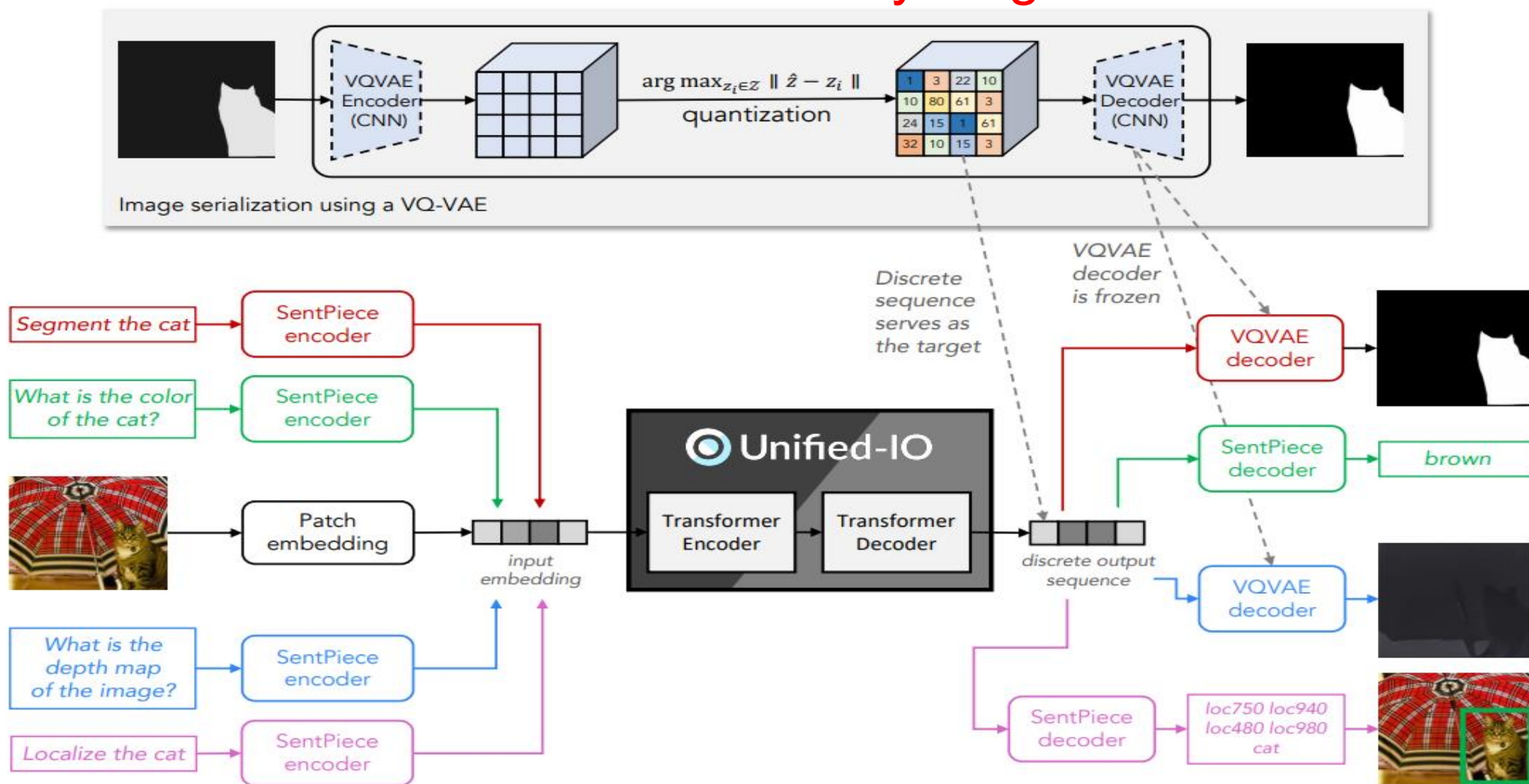


Figure 2: **Unified-IO**. A schematic of the model with four demonstrative tasks: object segmentation, visual question answering, depth estimation and object localization.

4 Models!

Model	Encoder Layers	Decoder Layers	Model Dims	MLP Dims	Heads	Total Params
UNIFIED-IO _{SMALL}	8	8	512	1024	6	71M
UNIFIED-IO _{BASE}	12	12	768	2048	12	241M
UNIFIED-IO _{LARGE}	24	24	1024	2816	16	776M
UNIFIED-IO _{XL}	24	24	2048	5120	32	2925M

Table 2: Size variant of UNIFIED-IO. Both encoder and decoder are based on T5 implementation ([Raffel et al., 2020](#)). Parameters of VQ-GAN ([Esser et al., 2021](#)) are not included in the total parameter count.

Datasets

Unified-IO Training Data

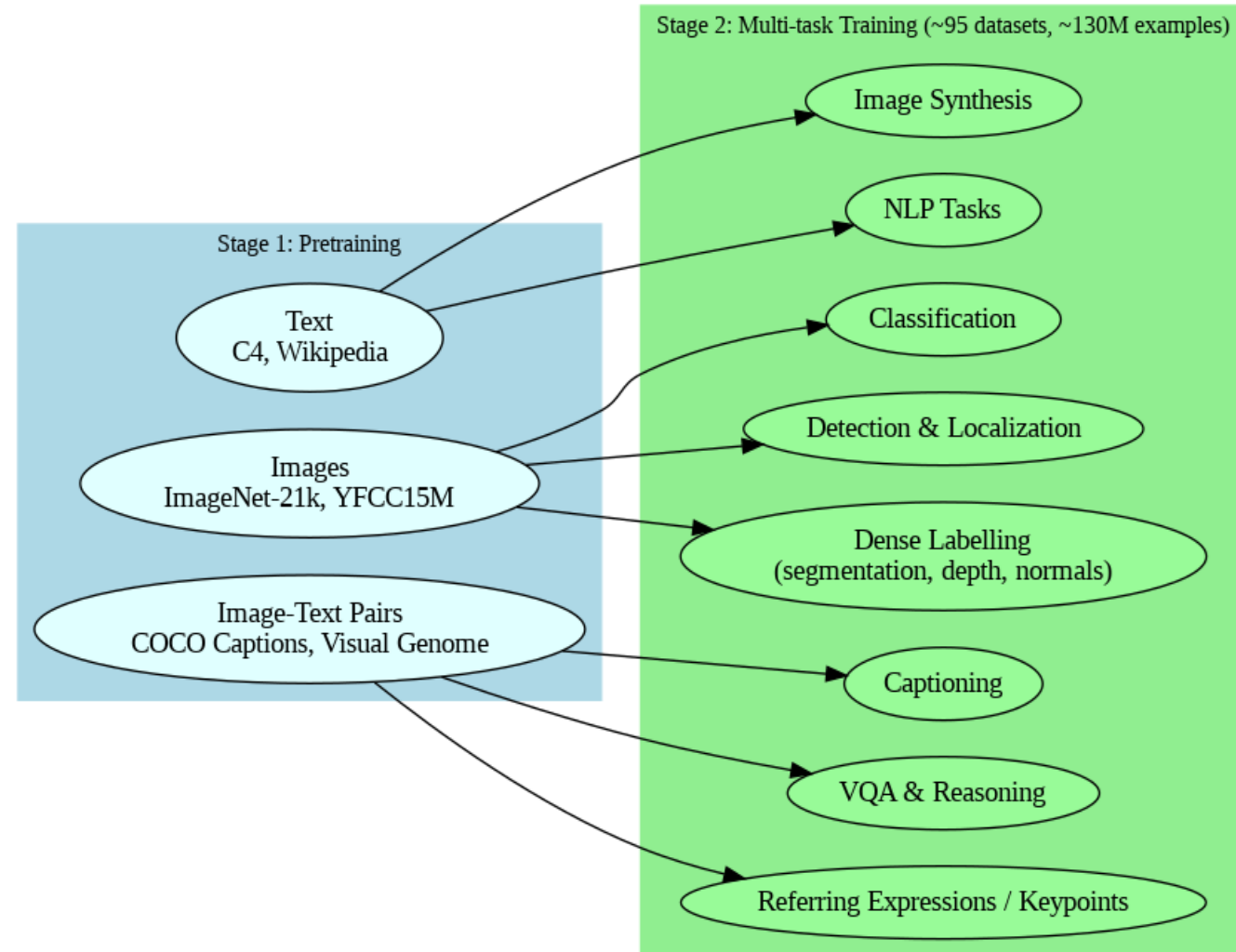
- Stage 1 (Pretraining):
- Text: C4, Wikipedia
- Images: ImageNet-21k, YFCC15M
- Image-Text pairs: COCO Captions, Visual Genome

Stage 2 (Multi-task training):

~95 datasets, grouped into 8 task categories:

- Image synthesis
- Detection & localization
- Dense labeling (segmentation, depth, normals)
- Captioning
- VQA & reasoning
- NLP tasks
- Classification
- Referring expressions / keypoints

Covers ~130M examples across all groups



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



What does the image describe?

a black bike parked next to a bed

Generate an image of "small personal pizza with bacon and spinach".



* Twenty-eight people were believed to have been spending Christmas Day with the caretaker of the St Sophia's camp, when the mudslide smashed into two cabins.

* Twenty-seven people were believed to have been spending Christmas Day with the caretaker of Saint Sophia Camp, a Greek Orthodox facility, when the mudslide roared through.

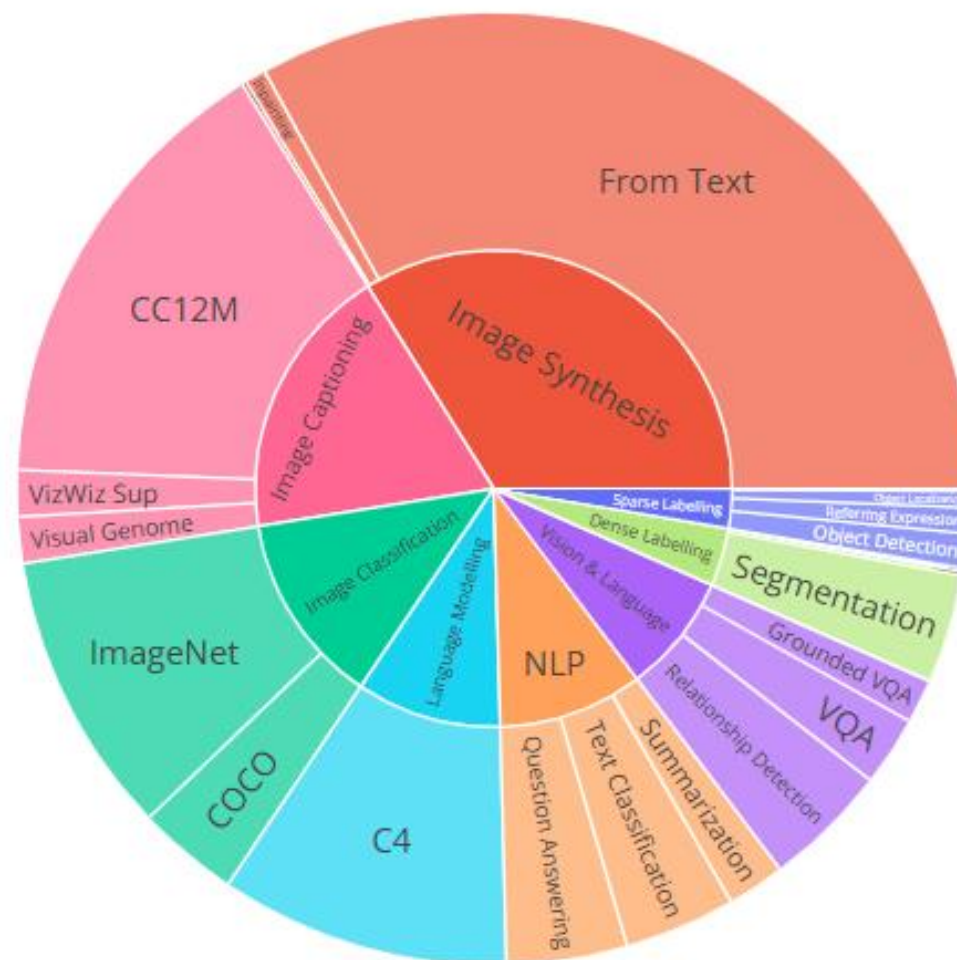
Yes, they are equivalent.

Training

Should multimodal models be trained from scratch or built on pretrained LLMs?

Two-Stage Training Pipeline

- Stage 1: Pretraining
 - Text denoising (mask 15% spans, reconstruct)
 - Image denoising (mask 75% patches, reconstruct via VQ-VAE tokens)
 - Sample datasets proportional to size
- Stage 2: Multi-task Joint Training
 - Train on all 95 datasets simultaneously
 - Sampling: balance groups (equal probability, except small adjustment for image synthesis/dense labeling)
 - Within groups: sample $\propto \sqrt{(\text{dataset size})}$ (so small datasets aren't drowned out)
- Models trained: 71M \rightarrow 2.9B parameters



		Example Source	Size				Input Modalities				Output Modalities			
			Datasets	Size	Percent	Rate	Text	Image	Sparse	Dense	Text	Image	Sparse	Dense
Image Synthesis			14	56m	43.0	18.7	✓	✓	✓	✓	-	✓	-	-
Image Synthesis from Text		RedCaps	9	55m	41.9	16.7	✓	-	-	-	-	✓	-	-
Image Inpainting		VG	3	1.2m	0.9	1.5	✓	✓	✓	-	-	✓	-	-
Image Synthesis from Seg.		LVIS	2	220k	0.2	0.6	✓	-	-	✓	-	✓	-	-
Sparse Labelling			10	8.2m	6.3	12.5	✓	✓	✓	-	-	-	✓	-
Object Detection		Open Images	3	1.9m	1.5	3.6	-	✓	-	-	-	-	✓	-
Object Localization		VG	3	6m	4.6	7.1	✓	✓	-	-	-	-	✓	-
Keypoint Estimation		COCO	1	140k	0.1	0.7	-	✓	✓	-	-	-	✓	-
Referring Expression		RefCoco	3	130k	0.1	1.1	✓	✓	-	-	-	-	✓	-
Dense Labelling			6	2.4m	1.8	6.2	✓	✓	-	-	-	-	-	✓
Depth Estimation		NYU Depth	1	48k	0.1	0.4	-	✓	-	-	-	-	-	✓
Surface Normal Estimation		Framenet	2	210k	0.2	1.1	-	✓	-	-	-	-	-	✓
Object Segmentation		LVIS	3	2.1m	1.6	4.7	✓	✓	-	-	-	-	-	✓
Image Classification			9	22m	16.8	12.5	-	✓	✓	-	✓	-	-	-
Image Classification		ImageNet	6	16m	12.2	8.1	✓	✓	-	-	✓	-	-	-
Object Categorization		COCO	3	6m	4.6	4.4	-	✓	✓	-	✓	-	-	-
Image Captioning			7	31m	23.7	12.5	-	✓	✓	-	✓	-	-	-
Webly Supervised Captioning		CC12M	3	26m	19.7	8.8	-	✓	-	-	✓	-	-	-
Supervised Captioning		VizWiz	3	1.4m	1.1	1.7	-	✓	-	-	✓	-	-	-
Region Captioning		VG	1	3.8m	2.9	2.0	-	✓	✓	-	✓	-	-	-
Vision & Language			16	4m	3.0	12.5	✓	✓	✓	-	✓	-	-	✓
Visual Question Answering		VQA 2.0	13	3.3m	2.5	10.4	✓	✓	✓	-	✓	-	-	-
Relationship Detection		VG	2	640k	0.5	1.9	-	✓	✓	-	✓	-	-	-
Grounded VQA		VizWiz	1	6.5k	0.1	0.1	✓	✓	-	-	✓	-	-	✓
NLP			31	7.1m	5.4	12.5	✓	-	-	-	✓	-	-	-
Text Classification		MNLI	17	1.6m	1.2	4.8	✓	-	-	-	✓	-	-	-
Question Answering		SQuAD	13	1.7m	1.3	5.2	✓	-	-	-	✓	-	-	-
Text Summarization		Gigaword	1	3.8m	2.9	2.5	✓	-	-	-	✓	-	-	-
Language Modelling			2	-	-	12.5	✓	-	-	-	✓	-	-	-
Masked Language Modelling		C4	2	-	-	12.5	✓	-	-	-	✓	-	-	-
All Tasks			95	130m	100	100	✓	✓	✓	✓	✓	✓	✓	✓



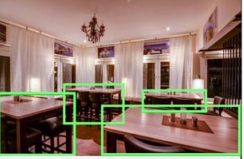







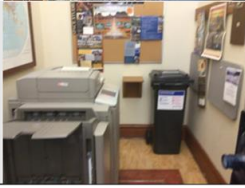
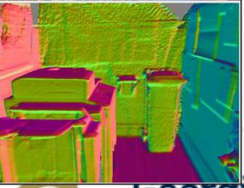
Results

GRIT benchmark: first model to do all 7 tasks, best score (64.3, +32 over prior best)

Performs well across 16 other benchmarks (ImageNet, VQA, NYU Depth, BoolQ, etc.)

Shows little drop from “seen” to “unseen” concepts → strong generalization

Not SOTA on every task, but competitive across board without fine-tuning

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			

Results

	<i>NYUv2</i>	<i>ImageNet</i>	<i>Place365</i>	<i>VQA_{v2}</i>	<i>OkVQA</i>	<i>A-OkVQA</i>	<i>VizWizQA</i>	<i>VizWizG</i>	<i>Swig</i>	<i>SNLI-VE</i>	<i>VisComet</i>	<i>Nocaps</i>	<i>COCO</i>	<i>COCO</i>	<i>MRPC</i>	<i>BoolQ</i>	<i>SciTail</i>
Split	val	val	val	test-dev	test	test	test-dev	test-std	test	val	val	val	val	test	val	val	test
Metric	RMSE	Acc.	Acc.	Acc.	Acc.	Acc.	Acc.	IOU	Acc.	Acc.	CIDEr	CIDEr	CIDEr	CIDEr	F1	Acc	Acc
Unified SOTA	UViM 0.467	- -	- -	- -	Flamingo 57.8	- -	Flamingo 49.8	- -	- -	- -	- -	- -	- -	- -	T5 92.20	PaLM 92.2	- -
UNIFIED-IO _{SMALL}	0.649	42.8	38.2	57.7	31.0	24.3	42.4	35.5	17.3	76.5	-	45.1	80.1	-	84.9	65.9	87.4
UNIFIED-IO _{BASE}	0.469	63.3	43.2	61.8	37.8	28.5	45.8	50.0	29.7	85.6	-	66.9	104.0	-	87.9	70.8	90.8
UNIFIED-IO _{LARGE}	0.402	71.8	50.5	67.8	42.7	33.4	47.7	54.7	40.4	86.1	-	87.2	117.5	-	87.5	73.1	93.1
UNIFIED-IO _{XL}	0.385	79.1	53.2	77.9	54.0	45.2	57.4	65.0	49.8	91.1	21.2	100.0	126.8	122.3	89.2	79.7	95.7
Single or fine-tuned SOTA	BinsFormer 0.330	CoCa 91.00	MAE 60.3	CoCa 82.3	KAT 54.4	GPV2 38.1	Flamingo 65.7	MAC-Caps 27.3	JSL 39.6	OFA 91.0	SVT 18.3	CoCa 122.4	- -	OFA 145.3	Turing NLR 93.8	ST-MOE 92.4	DeBERTa 97.7

When we evaluate models like Unified-IO, should we prioritize broad generalization across many tasks, or top performance on individual benchmarks?

Strengths

True unification of modalities and tasks

- One seq2seq Transformer handles 95 datasets / 22 tasks / 8 groups with no task-specific heads (Sec. 3.1, Fig. 2).
- Competitive across perception (detection, segmentation), generation (captioning, image synthesis), and reasoning (VQA, NLVR2).

Strong generalization across tasks

- On the GRIT benchmark, Unified-IO-XL is the only model that supports all 7 tasks and achieves the highest average (64.3 vs 32.0 for GPV-2) (Table 3).
- Maintains performance across “seen” vs “unseen” prompts and datasets (Sec. 5.3).

Scalable and flexible

- Model scales up to 2.9B parameters and shows consistent gains with size (Table 4).
- Outputs are always token sequences, so the same infrastructure can be extended to new modalities.

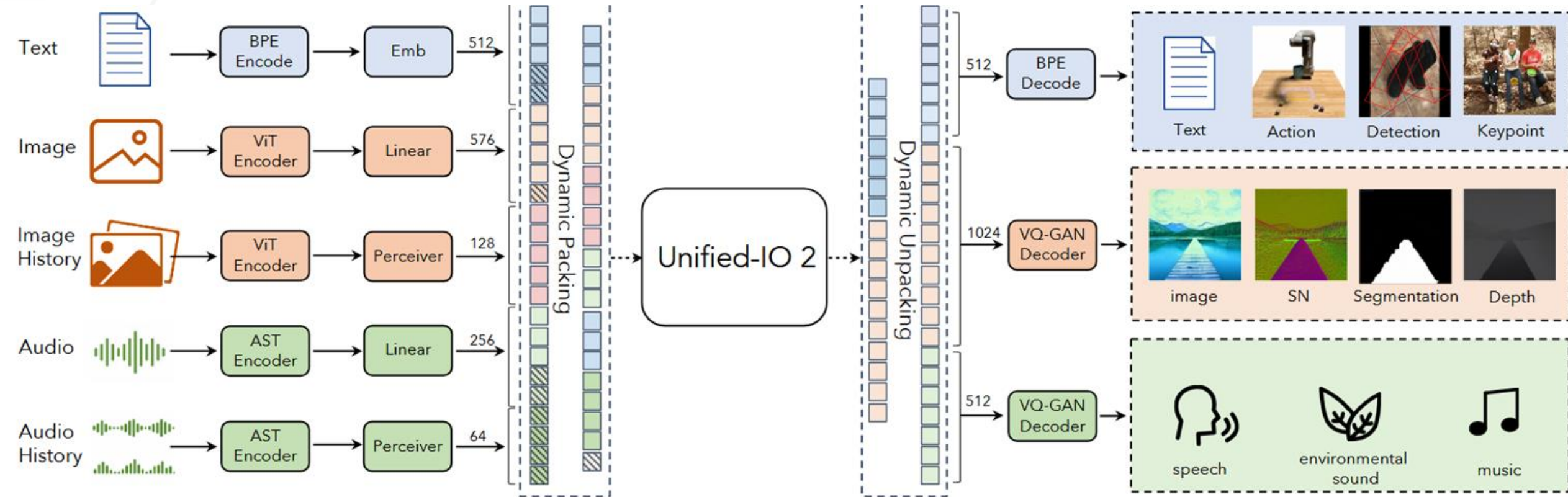
Simplified I/O representation

- Unified vocabulary (49,536 tokens: 32k text, 16k vision, 1k location) lets everything be cast as sequence prediction (Sec. 3.2).

Weaknesses

- **Detection struggles in cluttered scenes**
 - Paper notes low recall in dense environments — bounding box outputs often miss small or overlapping objects (Sec. 5.3, error analysis).
- **Image generation capped by VQ-VAE quality**
 - Frozen VQ-VAE used for image tokens → limits fidelity, produces blurrier generations compared to diffusion-based models (Sec. 3.2 + Appendix B).
- **Prompt sensitivity**
 - Case study on RefCOCO: small changes in prompt phrasing cause large accuracy drops (Table 7). Shows the model doesn't robustly generalize across linguistic variations.
- **Language weaker than vision**
 - Performs “respectably” on NLP tasks, but far below large LLMs trained on trillions of tokens (Sec. 5.2, Table 6).
This is a scale issue: max 2.9B params vs 100B+ for frontier LMs.
- **Task imbalance in training**
 - Even with $\sqrt{(\text{dataset size})}$ sampling, rare tasks (e.g., depth) were sampled only 0.43% of the time (Appendix C). Limits ceiling on specialized tasks.

UnifiedIO-2

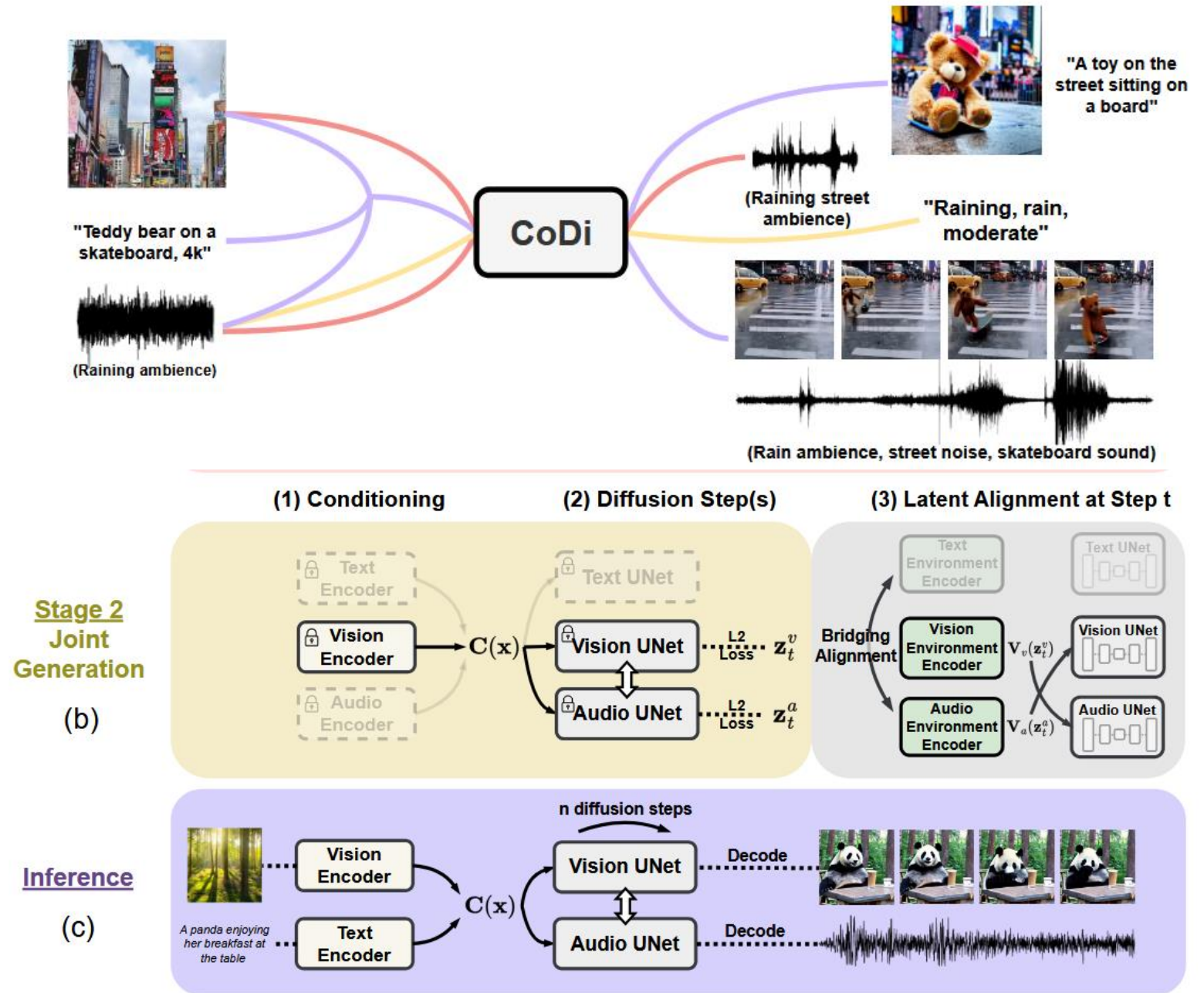


Now with audio, video, and action capabilities!

Related Works

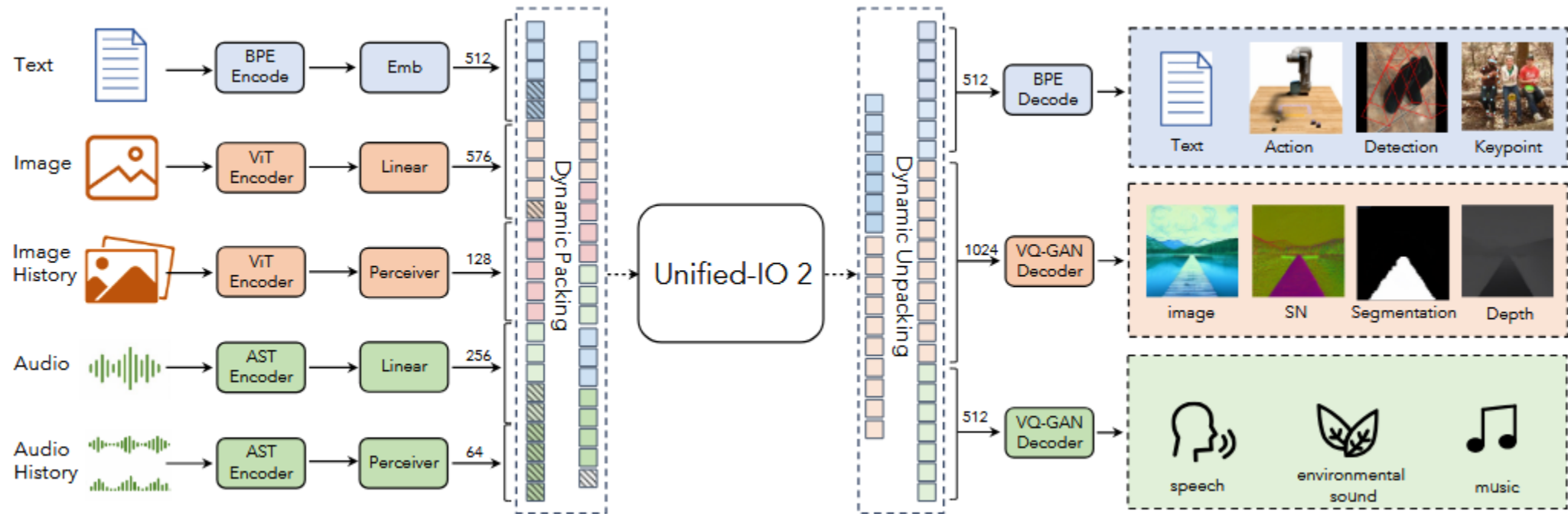
CoDi

Large diffusion layers decode an aligned latent space in its multi-headed architecture.



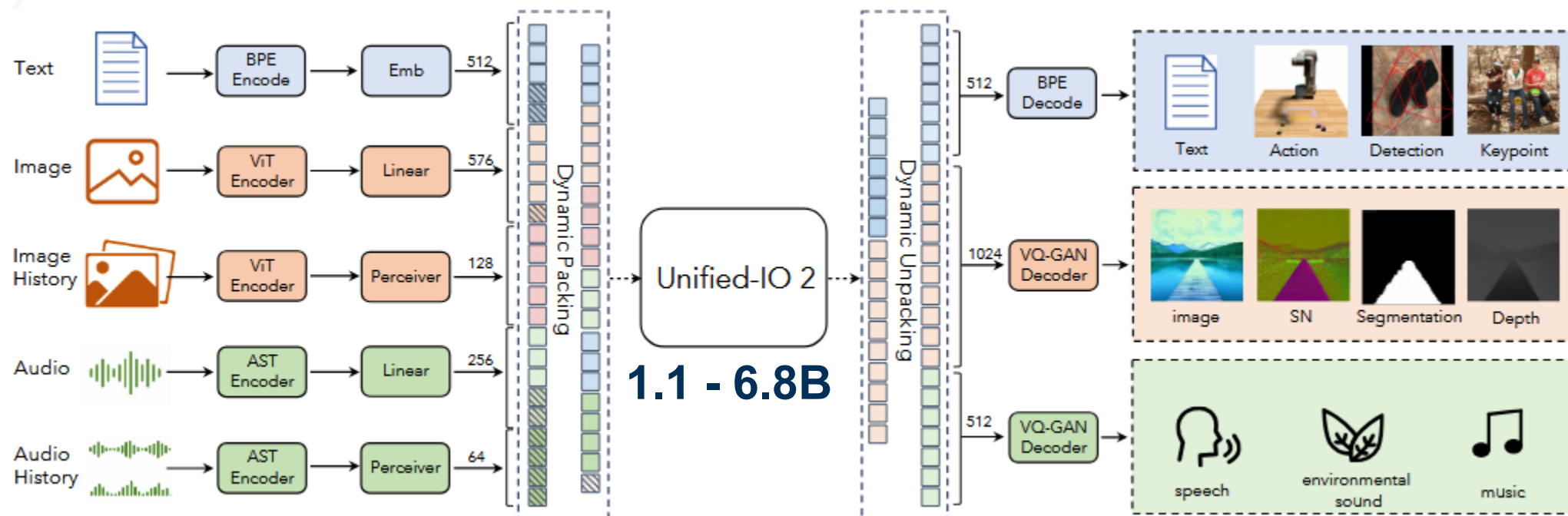
Motivation & Problem Statement

One autoregressive architecture to rule them all (text, vision, audio)

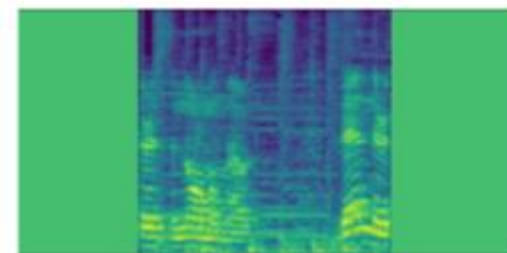


Rather than separate generator models, it uses **one** with thin decoding heads.

Architecture



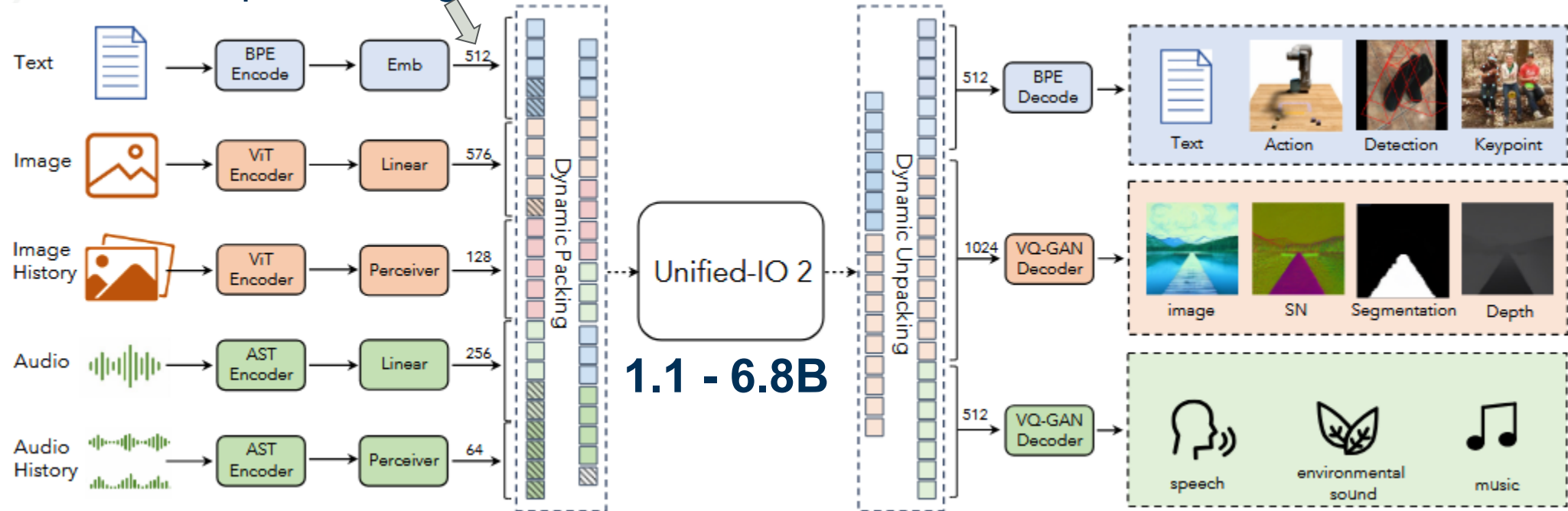
- Three input encoders
 - Text: LLaMA
 - Images: ViT
 - Audio: Audio Spectrogram Transformer



(Audio)

Architecture

Maximum sequence length

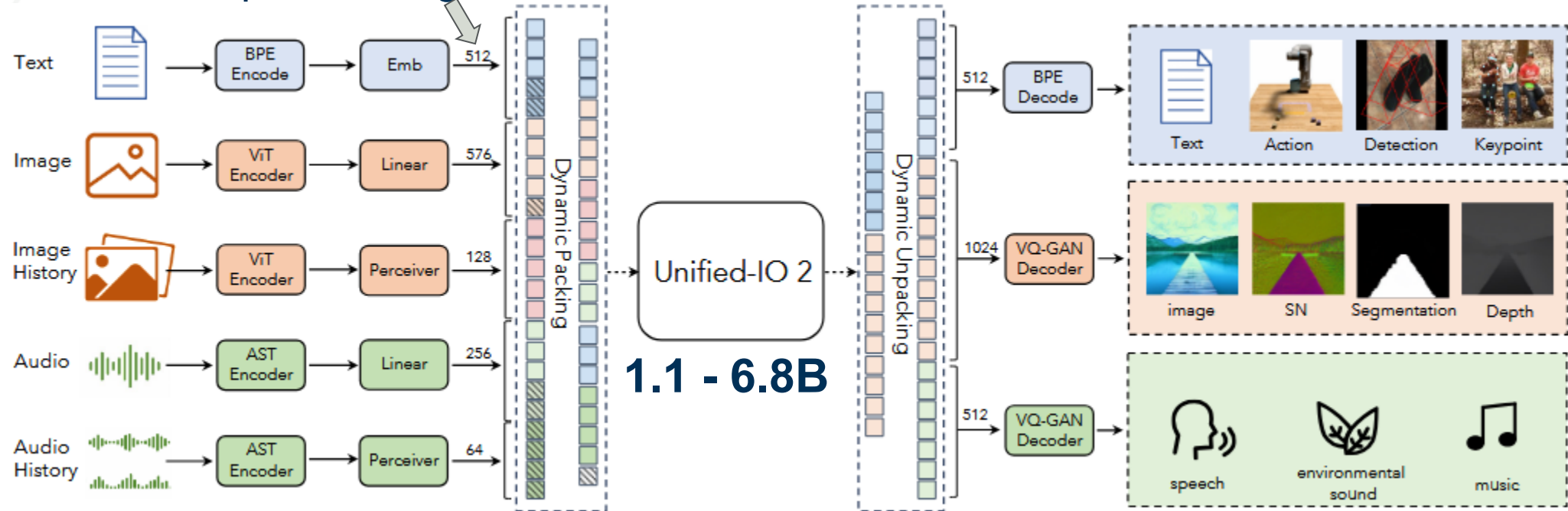


- Three input encoders
 - Text: LLaMA
 - Images: ViT
 - Audio: Audio Spectrogram Transformer
- 24 UIO-2 encoder x 24 decoder layers x (16 or 24) attention heads

Architecture

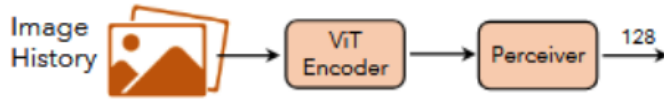
Do you agree with the sequence length budget allotment?

Maximum sequence length



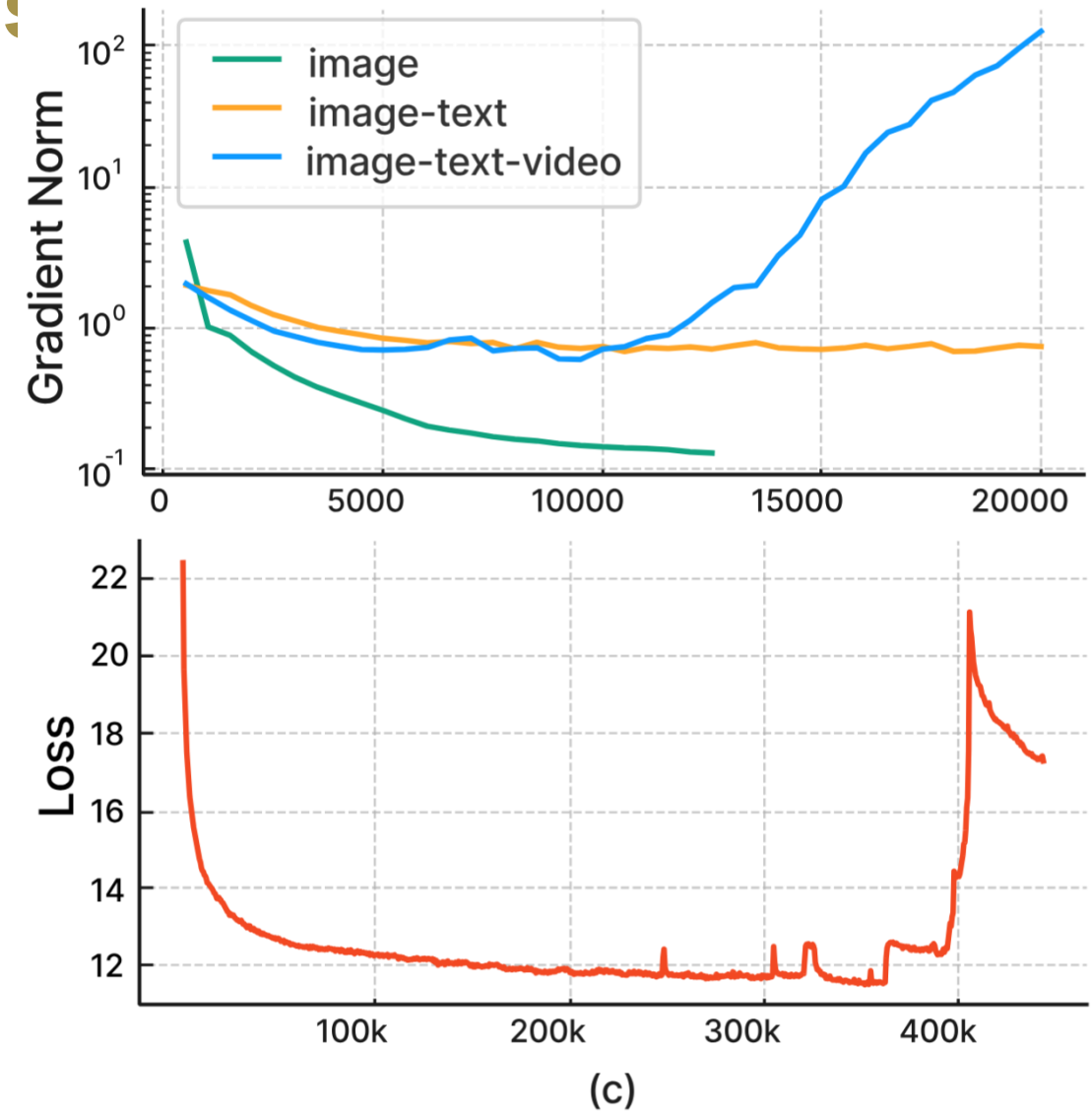
- Three input encoders
 - Text: LLaMA
 - Images: ViT
 - Audio: Audio Spectrogram Transformer
- 24 UIO-2 encoder x 24 decoder layers x (16 or 24) attention heads

Mo' modalities, mo' problems

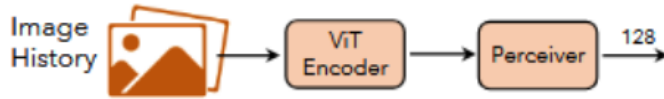


Gradient stability achieved through:

- **2D rotary encodings**
- **QK normalization**
- **Scaled cosine attention**
- Mixture of training
- Z-loss



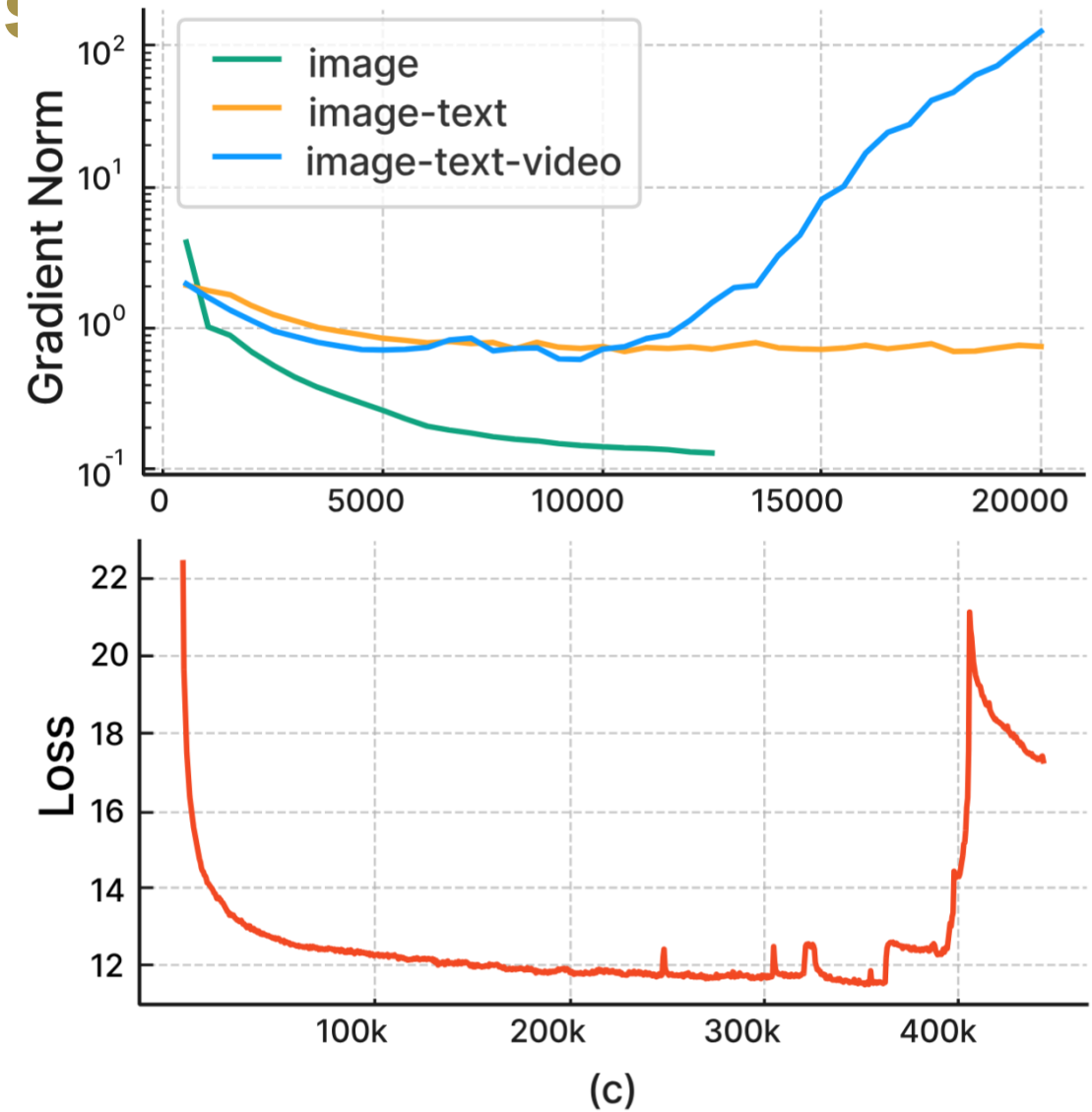
Mo' modalities, mo' problems



Gradient stability achieved through:

- **2D rotary encodings**
- **QK normalization**
- **Scaled cosine attention**
- Mixture of training
- Z-loss

$$\text{score}(i, j) = \frac{Q_i}{\|Q_i\|} \cdot \frac{K_j}{\|K_j\|} \cdot \alpha$$

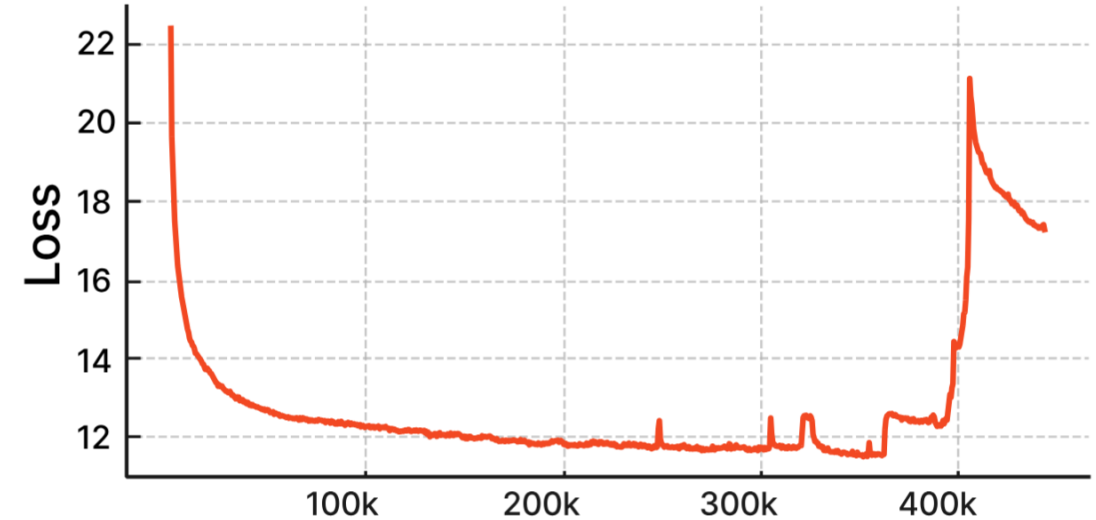
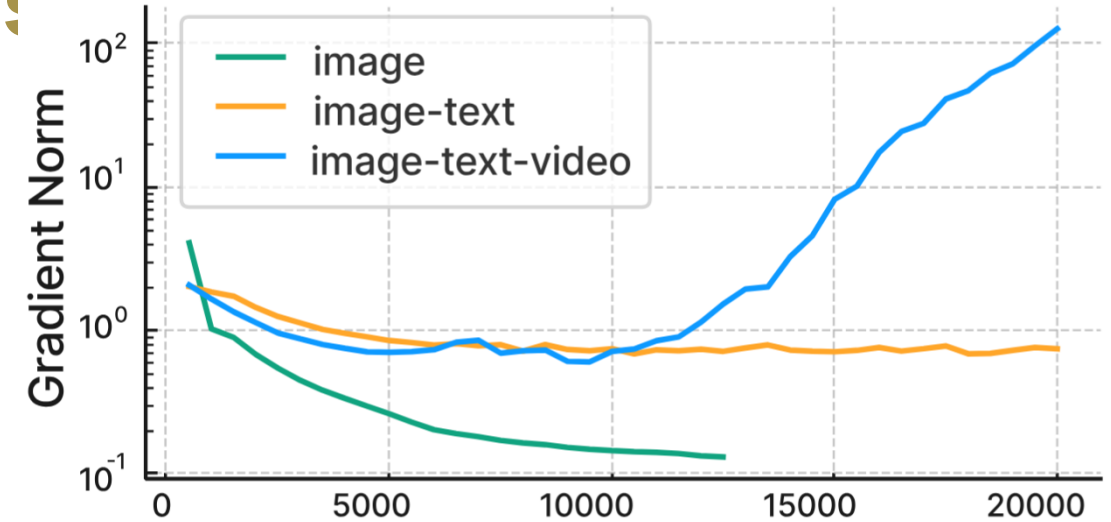


Mo' modalities, mo' problems

Gradient stability achieved through:

- **2D rotary encodings**
- **QK normalization**
- **Scaled cosine attention**
- Mixture of training
- Z-loss

$$Z = \sum_j e^{z_j}$$
$$\mathcal{L}_Z = \mathcal{L}_{CE} + \lambda \cdot (\log Z)^2$$

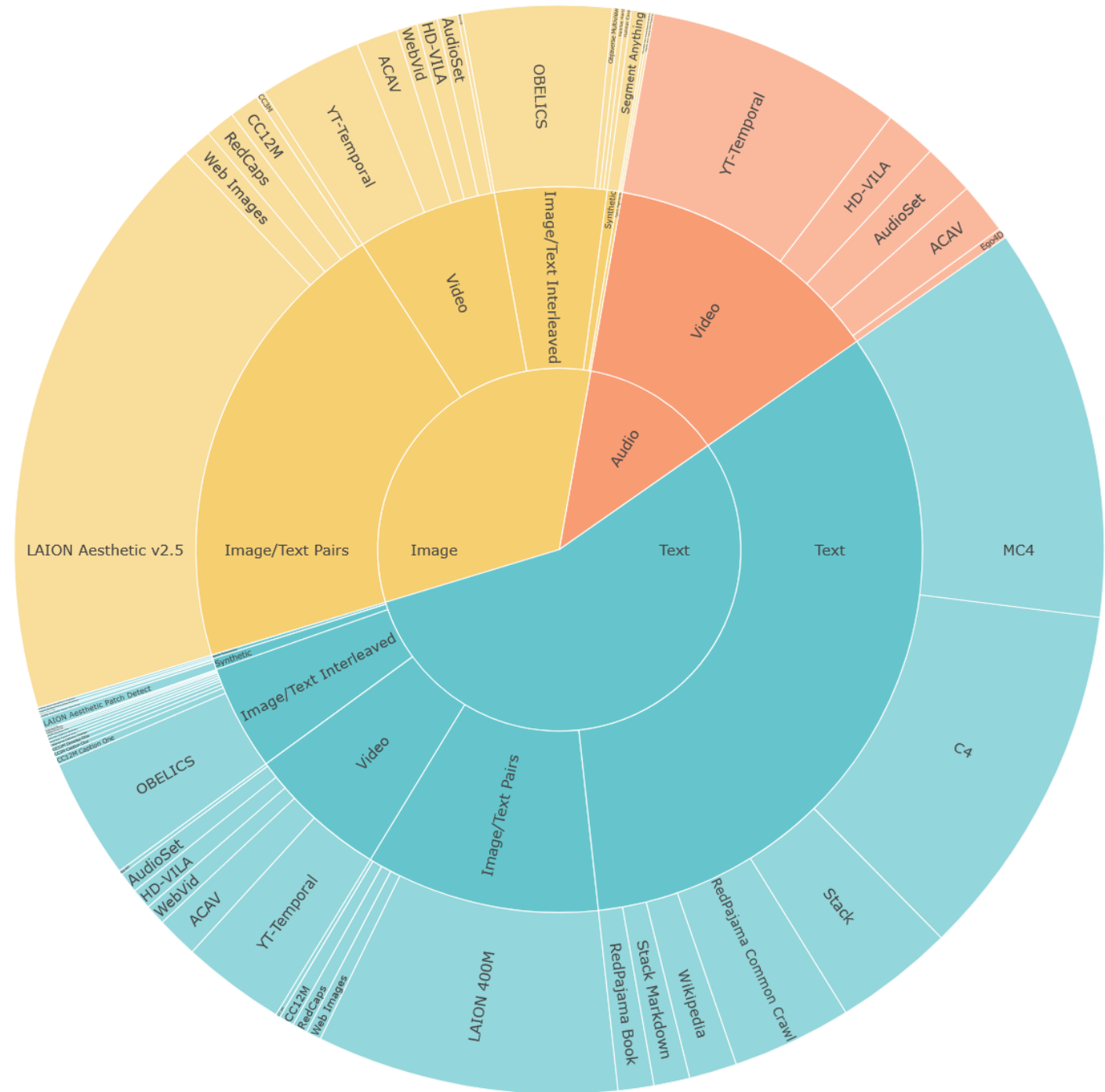


(c)

Pre-training Data

600 TB

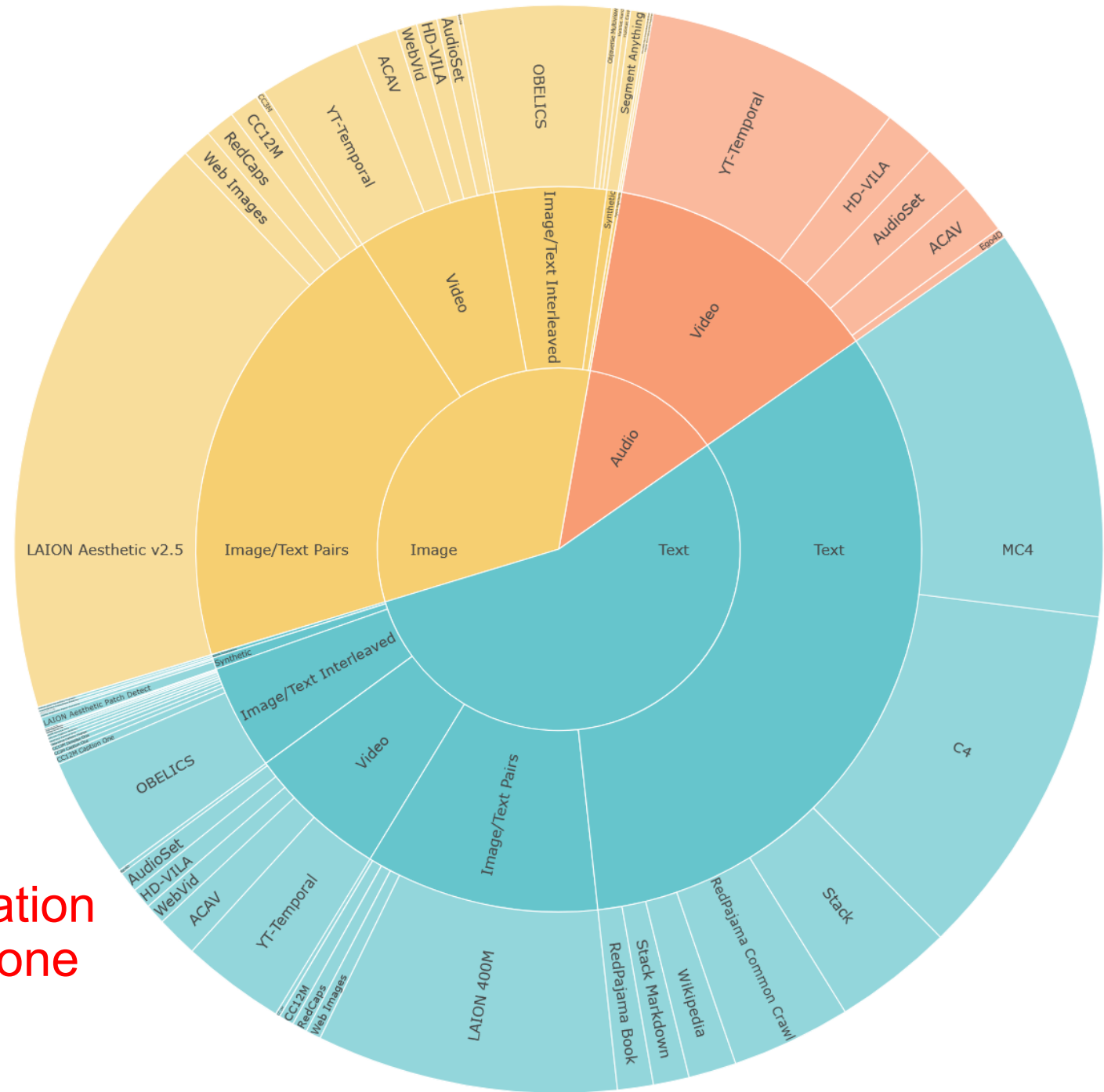
- NLP [33%]
- Image & Text Pairs [40%]
- A/V [25%]
- 3D Embodiment [1%]
- + Instruction fine tuning



Pre-training Data

600 TB

- NLP [33%]
- Image & Text Pairs [40%]
- A/V [25%]
- 3D Embodiment [1%]
- + Instruction fine tuning



Embeddings share a single representation space, how well represented can any one modality be?

Instruction Tuning Dat

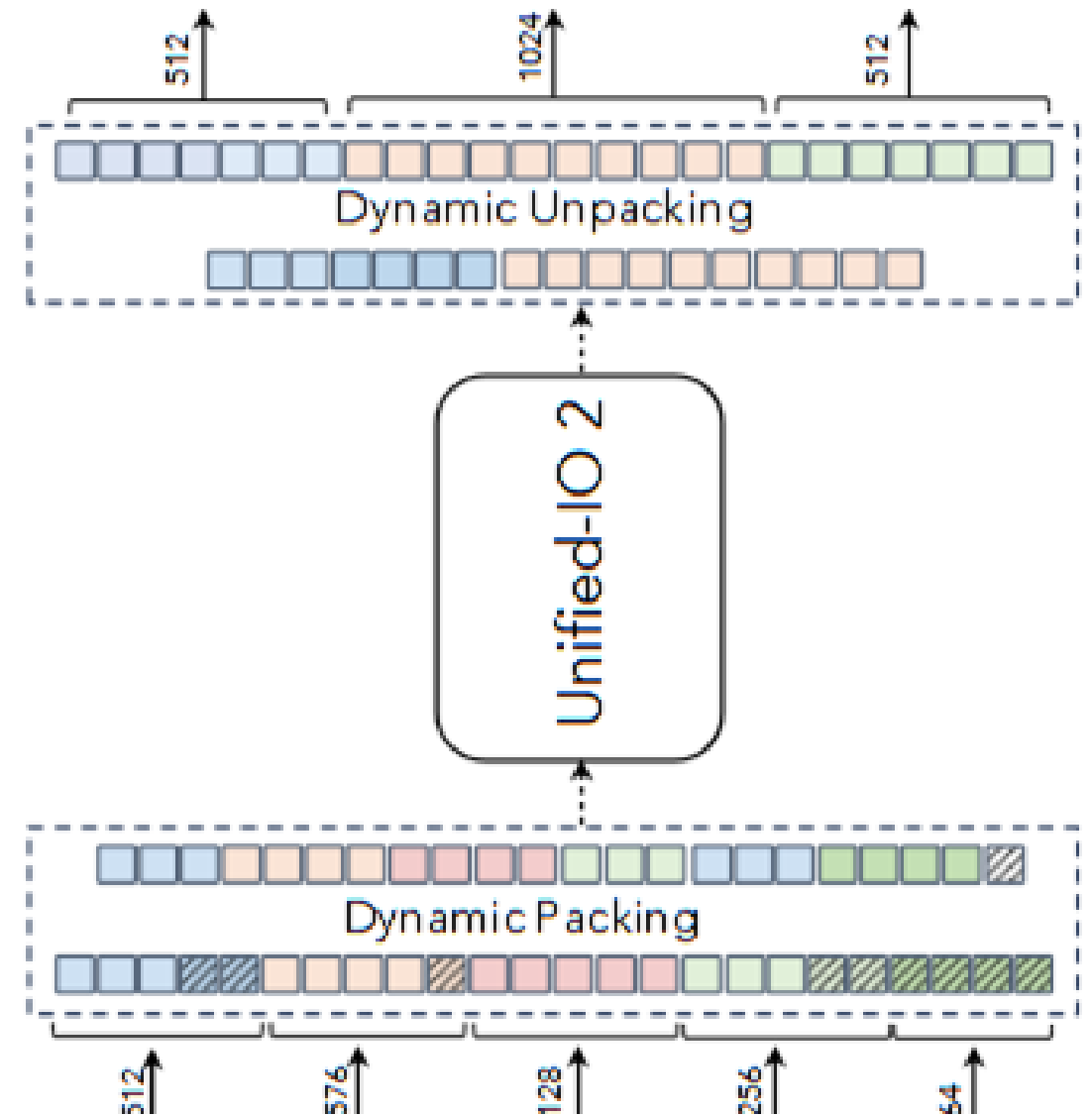
- **Language [25%]**
- **Image [49%]**
 - Generation [17.6%]
 - Reasoning [17.8%]
 - Sparse image labelling [7.25%]
 - Surface normal & depth estimation [4.1%]
- **Audio [18%]**
 - Generation [7.5%]
 - Captioning [10.6%]
- **Video [14%]**
 - Reasoning [10.6%]
 - Sparse Labelling [3.42%]
- **Embodied AI [4%]**



Training - Dynamic Packing

1. Encode training examples in batch
2. Concatenate examples up to 1152 token limit
3. Mask attention layers to attend to only their example
4. Unpack into a maximum of 2048 tokens

Most training examples only include a **handful of modalities**, so we process **multiple training examples**

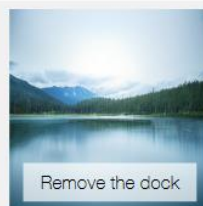


Capabilities

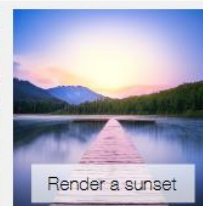
Image Editing



Remove the dock

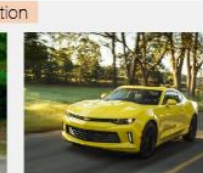


Paint this image like Van Gogh



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:

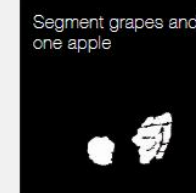
Ingredients:
- 1 cup all-purpose flour, - ½ cup sugar ...
Instructions:
1. In a large bowl whisk together the flour, sugar..
2. In a separate bowl, mix together the eggs ...

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.



Visual based Audio Generation



Generate an audio track for this band.



Generate an audio track for this band.



Visual Audio Localization



Identify the locations of the instruments producing the given sound.



Robotic Manipulation

= blicket
 = dax
 is kobar than
Put the less kobar blicket into the dax.



Future Frame Prediction



Given the initial image and a sequence of actions, predict the next frames

Generated Images



Capabilities

Visual based Audio Generation



Generate an audio track for this band.



Generate an audio track for this band.

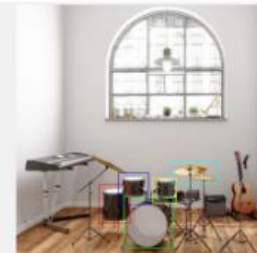


Visual Audio Localization







(drum sounds)

Identify the locations of the instruments producing the given sound.



Robotic Manipulation

 = blicket
 = dax
 is kobar than 
Put the less kobar blicket into the dax.



Future Frame Prediction



Given the initial image and a sequence of actions, predict the next frames

Generated Images



Results - Pre-instruction Tuning

re-instruction Tuning

	NLP	Text & Image	Spatial-Causal Reasoning	Audio Reasoning	
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

Results - Generation

Realism
Text/Image
Alignment

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

Results - Vision/Lang Reasoning

Method	VQA ^{v2}	OKVQA	SQA	SQA ¹	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

Audio-Video Reasoning

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

Ablation Study

	AP3D	AP3D@15	AP3D@25	AP3D@50
Cube-RCNN [16]	50.8	65.7	54.0	22.5
UIO-2 _L	42.9	54.4	45.7	21.7
UIO-2 _{XL}	43.3	54.4	46.8	21.8
UIO-2 _{XXL}	42.4	54.0	45.6	20.9

Table 7. Single-object 3D detection results on Objectron [3].

Grounding



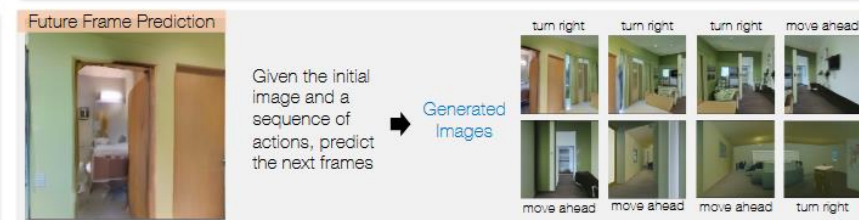
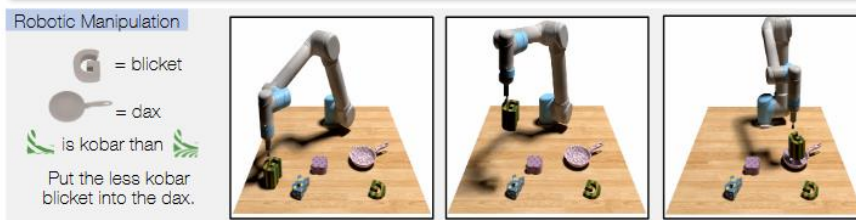
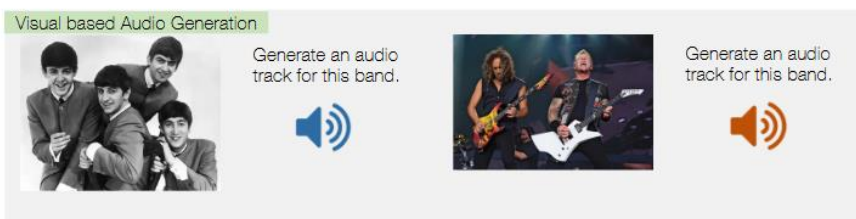
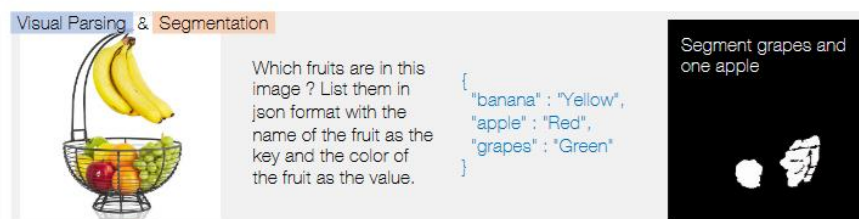
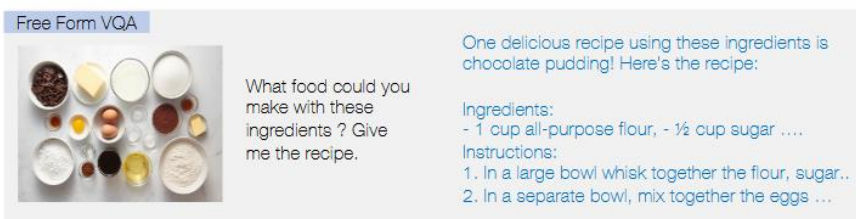
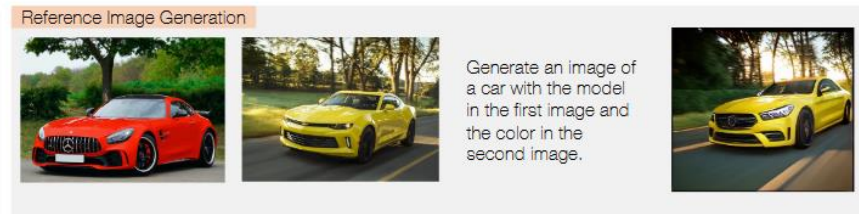
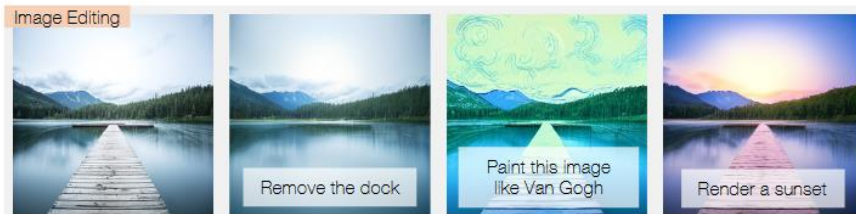
	Method	Cat.	Loc.	Vqa	Ref.	Seg.	KP	Norm.	All
Ablation	UIO-2 _L	70.1	66.1	67.6	66.6	53.8	56.8	44.5	60.8
	UIO-2 _{XL}	74.2	69.1	69.0	71.9	57.3	68.2	46.7	65.2
	UIO-2 _{XXL}	74.9	70.3	71.3	75.5	58.2	72.8	45.2	66.9
Test	GPV-2 [89]	55.1	53.6	63.2	52.1	-	-	-	-
	UIO _{XL} [123]	60.8	67.1	74.5	78.9	56.5	67.7	44.3	64.3
	UIO-2 _{XXL}	75.2	70.2	71.1	75.5	58.8	73.2	44.7	67.0

Table 3. Results on the GRIT ablation and test sets [66].

Why do you think performance dropped between UIO-1 and 2?

Conclusions

The first
autoregressive multi-
modal model that does
vision, text, audio,
and **action.**



But should it be the last?

Small Language Models are the Future of Agentic AI

Peter Belcak, Greg Heinrich, Yonggan Fu, Xin Dong, Saurav Muralidharan, Yingyan Celine Lin,
Pavlo Molchanov
NVIDIA Research

- **Leverage SLMs for Rapid Specialization.** Teams should take advantage of the agility of SLMs by fine-tuning them for specific tasks, enabling faster iteration cycles and easier adaptation to evolving use cases and requirements.

Thank you!

Questions or
Thoughts?

Three training modes

