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

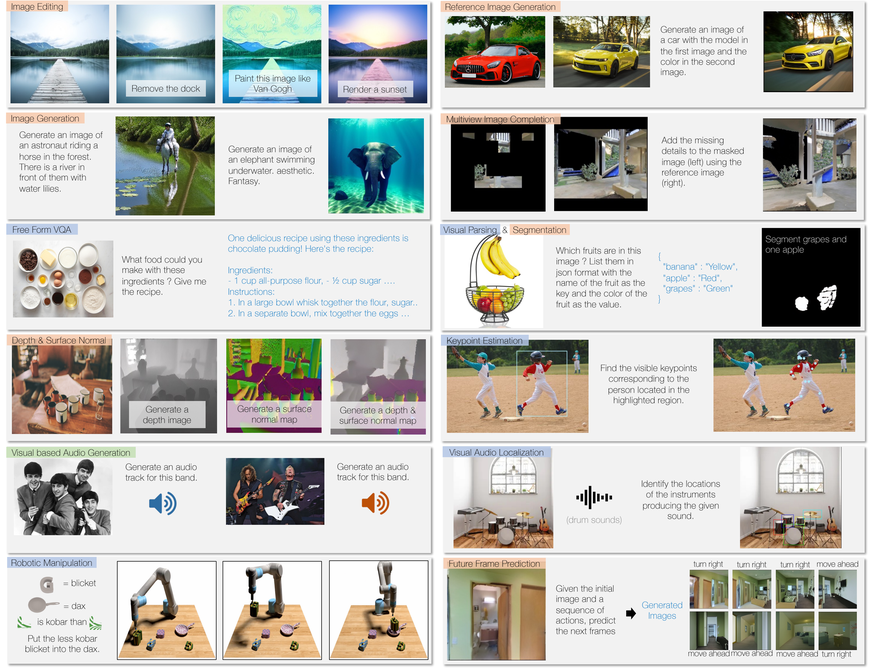
Jiasen Lu*, Christopher Clark*, Sanghoo Lee*, Zichen Zhang*, (*Leading Authors, equal contribution.)
Savya Khosla, Ryan Marten, Derek Hoiem , Aniruddha Kembhavi

Allen Institute for AI
ILLINOIS UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

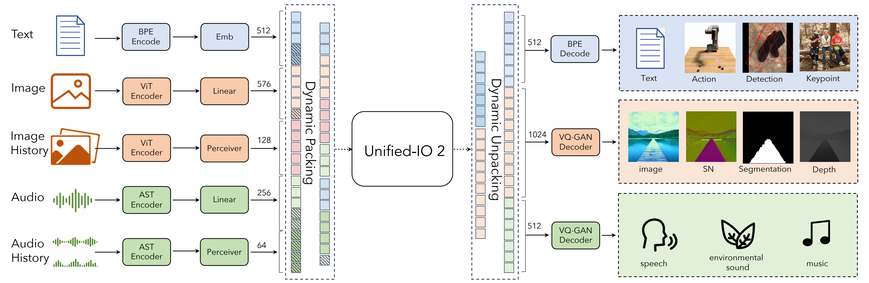
CVPR JUNE 17-21, 2024 SEATTLE, WA

Code + Model!

1 Early-Fusion token-based Mixed-modal models

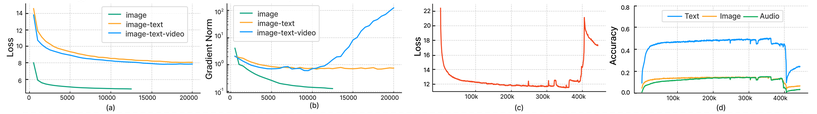


2 Model Architecture



3 Training Recipes

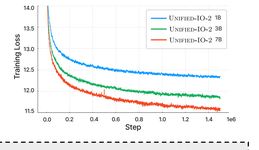
Training fails with native LLM recipes



(a)(b): Gradient norm grows as adding more modalities. (c)(d): Loss explodes after 350k steps.

Reason: numerical instabilities caused by multi-modal

- QK-Normalization
- Float32 logit attention
- Freeze ViT and AST
- Z-loss
- 2D Rotary Embedding
- Scaled Cosine Attention
- Mixture of Training
- …

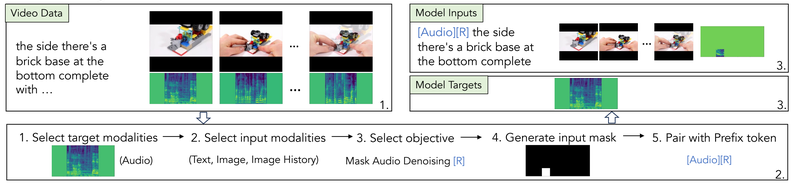


Pre-Training Objective: Multimodal Mixture of Denoiser

[R] – standard span corruption
[S] – causal language modeling
[X] – extreme span corruption
[R] – masked denoising of input image or audio patches
[S] – generate target conditioned on other input modalities
Modality tokens – [Image], [Text], [Audio]

4 Training Data

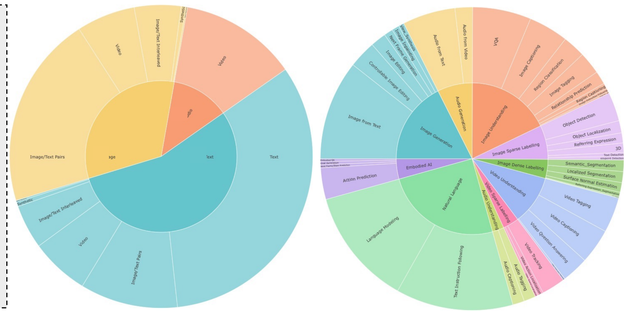
Pre-Training Data Conversion



Pre-Training and Instruction tuning data distribution

Pre-Training
- NLP [33%]
- Image & Text [40%]
- Video & Audio [25%]
- 3D & Embodiment [1%]
- Augmentation [1%]

Instruction Tuning
- ~800 million examples
- ~230 tasks
- Sample tasks balanced by modalities and dataset size



5 Main Results

- Wide range of tasks across modalities.
- ***Single*** model for all evaluation tasks.
- Zero-shot language similar to 3B LLM.
- Sota performance on GRIT, V&L.
- Competitive performance on various of tasks.

ipad demo region

Zero-shot performance

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 benchmark

	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**

Vision/Language

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$^{\diamond}$	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$^{\diamond}$	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$^{\diamond}$	125.4	87.7	61.8	71.5

Image/Audio/Action Generation

Method	Image FID↓	Image TIFA↑	Audio FAD↓	Audio IS↑	Audio KL↓	Action 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

Video/Audio Understanding

Method	Video: Kinetics-400 [90]	VATEXCaption [166]	MSR-VTT [169]	MSRVTT-QA [198]	MSVD-QA [198]	STAR [196]	SEED-T [166]	Audio: 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

Acknowledgement: This research was made possible with cloud TPUs from Google's TPU Research Cloud (TRC).