

Unified-IO 2

Levi Harris, Xinyu Liu

Background

- LLMs (e.g., ChatGPT) become powerful chatbots using instruction tuning
 - Text-only models
- LMMs (e.g, GPT-4V) extend LLM capabilities to *many modalities*
 - Images, videos, etc.
 - Can solve tasks across *many domains*
- **Problem:** more modalities + more data = complex models



GPT-3.5
Text In
Text Out



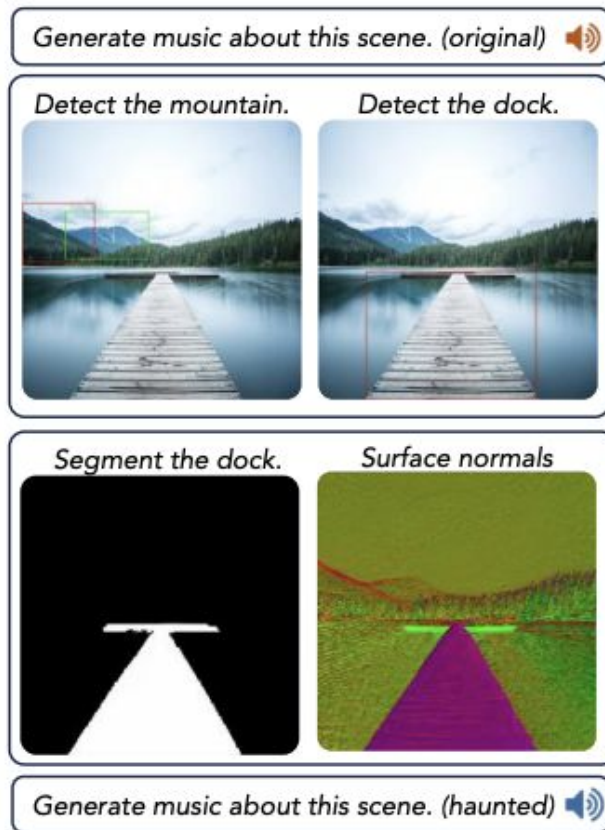
GPT-4V
Text + Images In
Text + Images Out*
**With DallEs*

+

Vision

Motivation

- Previous Models
 - Used *pre-trained LLMs*
 - Multiple models
 - Lack generative capabilities
 - Closed source
 - Complex
- Unified Backbone
 - Leverage data redundancies
 - Learn shared representations
 - Create an *anything in, anything out* assistant



*Unified-IO is robust to many
tasks and modalities*

Overview

- UNIFIED-IO 2 processes all modalities with a single, unified encoder-decoder transformer

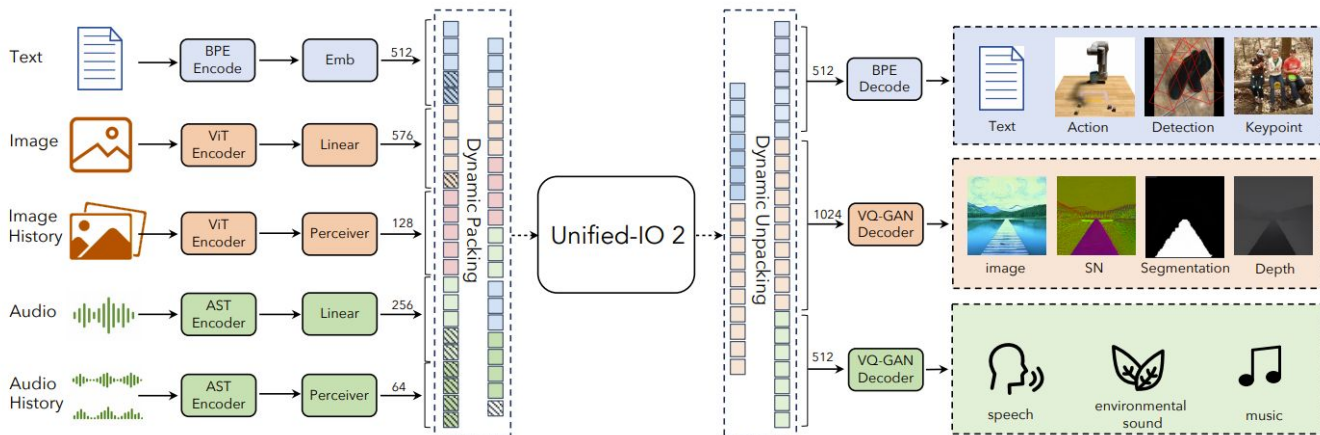


Figure 2. UNIFIED-IO 2 architecture. Input text, images, audio, or image/audio history are encoded into sequences of embeddings which are concatenated and used as input to an encoder-decoder transformer model. The transformer outputs discrete tokens that can be decoded into text, an image, or an audio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

Unified Task Representation

	Encode	Generate
Text, Sparse Structures, and Action	Text: the byte-pair encoding (LLaMA) Sparse Structures: 1000 special tokens Robotic Action: text commands + special tokens	the byte-pair decoder
Images and Dense Structures	a pre-trained ViT (feature from the second and second-to-last layers) + a linear layer	VQ-GAN model with 8×8 patch size that encodes a 256×256 image into 1024 tokens with a codebook size of 16512
Audio	spectrogram -> a pre-trained Audio Spectrogram Transformer (AST) + a linear layer	ViT-VQGAN with 8×8 patch size that encodes a 256×128 spectrogram into 512 tokens with a codebook size of 8196
Image and Audio History	the ViT/AST + a perceiver resampler	/

Unstable Training

- Using a standard implementation following UNIFIED-IO leads to increasingly unstable training as we integrate additional modalities.

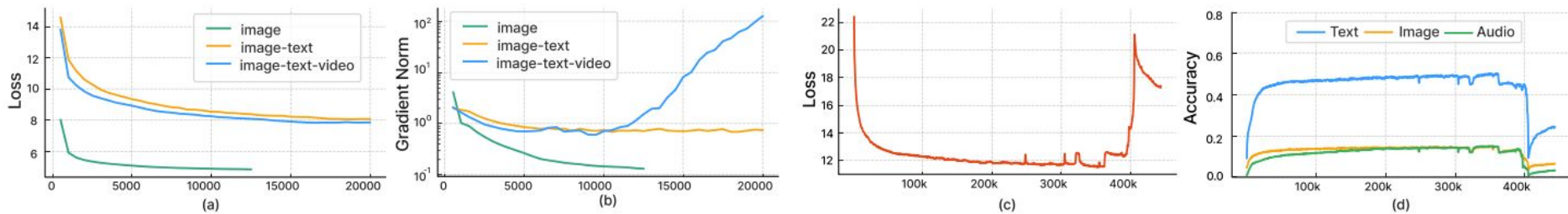


Figure 3. **Left:** Training loss (a) and gradient norms (b) on different modality mixtures. **Right:** Training loss (c) and next token prediction accuracy (d) of UIO-2_{XXL} on all modalities. Results were obtained before applying the proposed architectural improvements.

Architectural Modifications

- 2D Rotary Embedding
- QK Normalization
- Scaled Cosine Attention
 - perceiver resampler

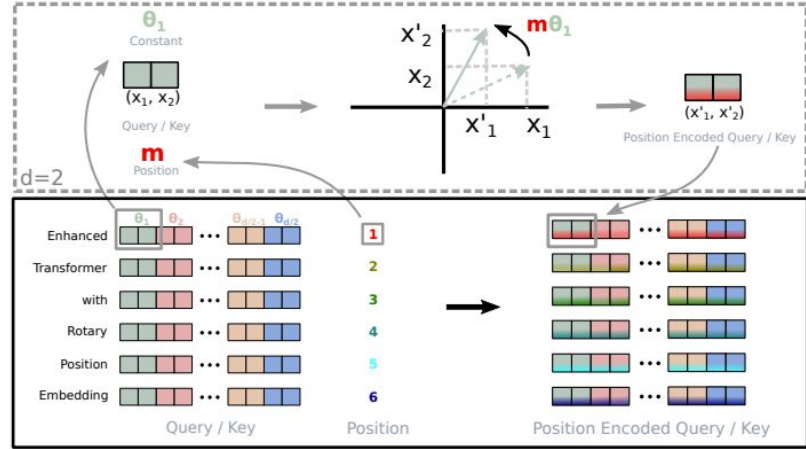


Figure 1: Implementation of Rotary Position Embedding(RoPE).

Training Objective

- Multimodal Mixture of Denoisers
 - Text
 - [R] – standard span corruption
 - [S] – causal language modeling
 - [X] – extreme span corruption
 - Image & Audio
 - [R] – masked denoising where $x\%$ of the input is masked and requires re-construction
 - [S] – generate the target modality conditioned only on other input modalities.

Training Objective

- Issue: information leak
- Autoregressive with Dynamic Masking
 - Mask the token in the decoder except when predicting that token

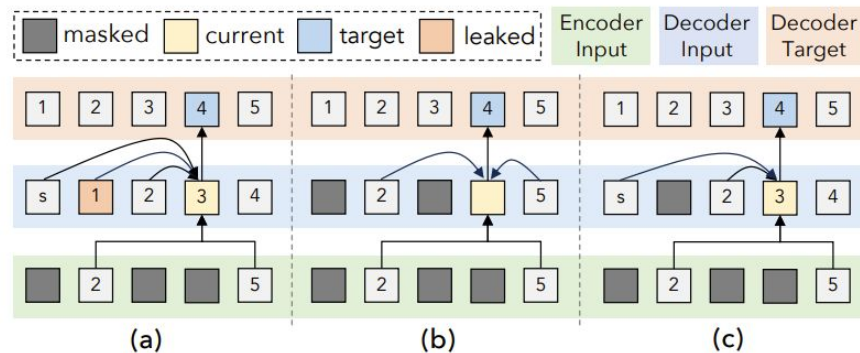


Figure 5. Different training paradigms in masked image modeling (a) autoregressive, (b) mask auto-encoder, (c) autoregressive with dynamic masking. Our proposed paradigms can maintain causal generation while avoiding information leaks in the decoder.

Efficient Implementation

- Issue: Heavily multimodal data -> highly variable sequence lengths
- Solution: Packing
 - Tokens of multiple examples are packed into a single sequence
 - The attentions are masked to prevent cross-attending between examples
- Optimizer: Adafactor

Multimodal Data

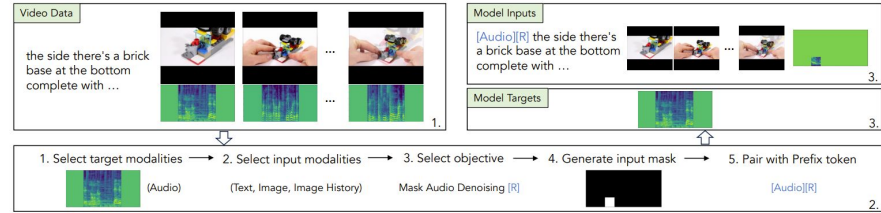


Figure 7. Construction of training samples from video data for the model's input and target. Given the video, we first extract the video frames and the corresponding audio spectrograms and transcript. Then, the data pass through a random selection process to determine the target modality, input modalities,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

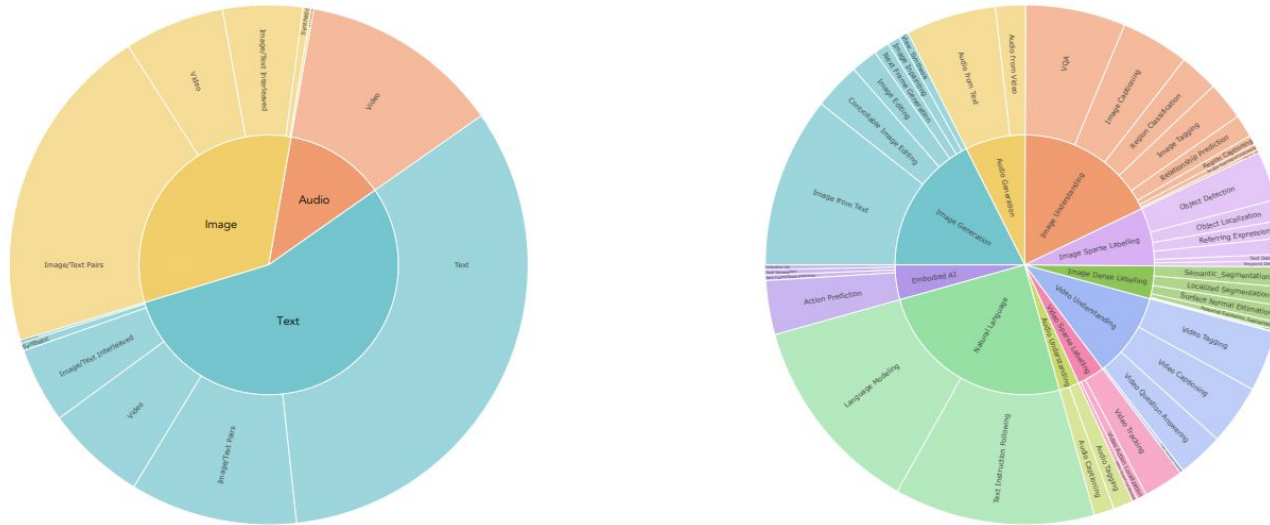


Figure 6. Distribution of pre-training and instruction tuning data. Segments proportional to sampling rates. The inner section shows the target modality, and the outer section shows the data type. Please refer to Figure 9 and Figure 11 in the Appendix for particular datasets.

Experimental Results

- Pre-Training Benchmarks
 - **Weak** performance on *language modeling tasks*
 - **Strong** performance on *text-to-image / text-to-audio* generation

- GRIT: Sparse + Dense Pixel

Prediction

- New **SOTA**
- Unified-IO 2 is a powerful, *general-purpose* model for language-image tasks

Method	HellaSwag \uparrow	TIFA \uparrow	SEED-S \uparrow	SEED-T \uparrow	AudioCaps \downarrow
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

Zero-shot pre-training benchmarks.

	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

Qualitative Results

<https://unified-io-2.allenai.org/>

Natural Language

User

“Give me a recipe for scrambled eggs.”



Model

“Scrambled eggs are a delicious and easy-to-make breakfast or brunch dish. Here is a simple recipe for scrambled eggs:

Ingredients:

- 2 large eggs...”

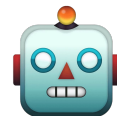


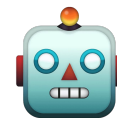
Image Generation

User

“A picture of a dog posing happily at landmarks on a trip to New York..”



Model



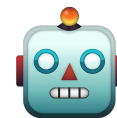
Audio Generation

User

“**Speak:** Printing, then, for our purpose, may be considered as the art of making books by means of movable types.”



Model



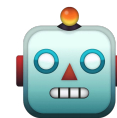
Audio Generation

User

“What is the
sound of this
instrument?”



Model



Video Understanding

User

“What is the robot doing?”



Model

“Move the orange object from the table to the toaster oven.”

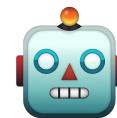
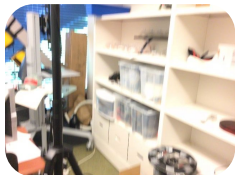


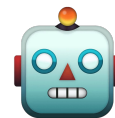
Image Dense Labeling

User

“Compute the surface normals for this image by encoding the orientation of each pixel into an image. Please use **red** for **z** orientation, **green** for **y** orientation.”

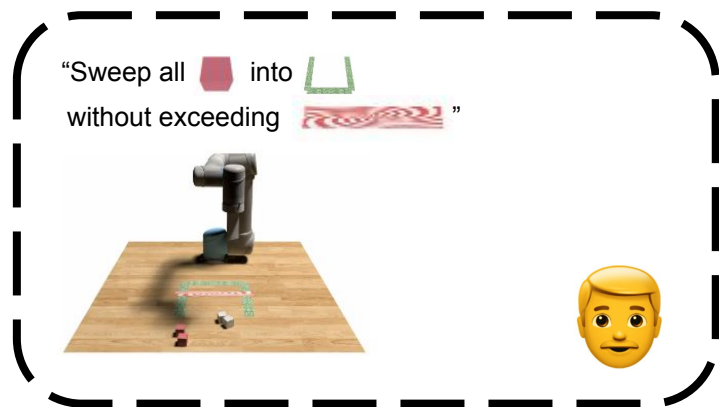


Model

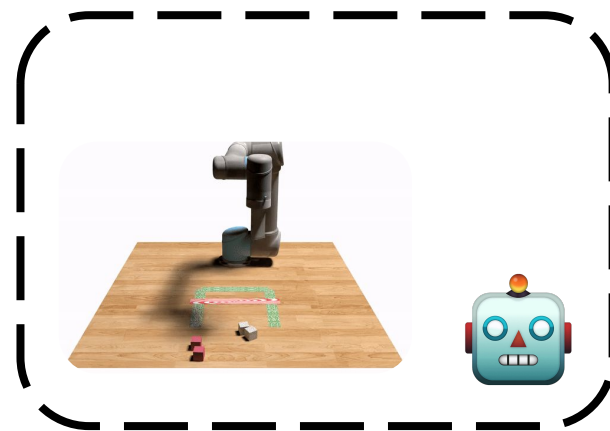


Embodied AI & 3D

User



Model



And much, much more!

Image Editing

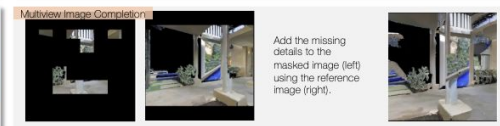
Reference Image Generation

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



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.

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

Questions?