UNIFIED-IO: A UNIFIED MODEL FOR VISION, LANGUAGE, AND MULTI-MODAL TASKS

Jiasen Lu[†,*] **Christopher Clark**[†*], **Rowan Zellers**[†◇],
Roozbeh Mottaghi[†◇], **Aniruddha Kembhavi**[†◇]
[†]Allen Institute for AI, [◇]University of Washington, Seattle

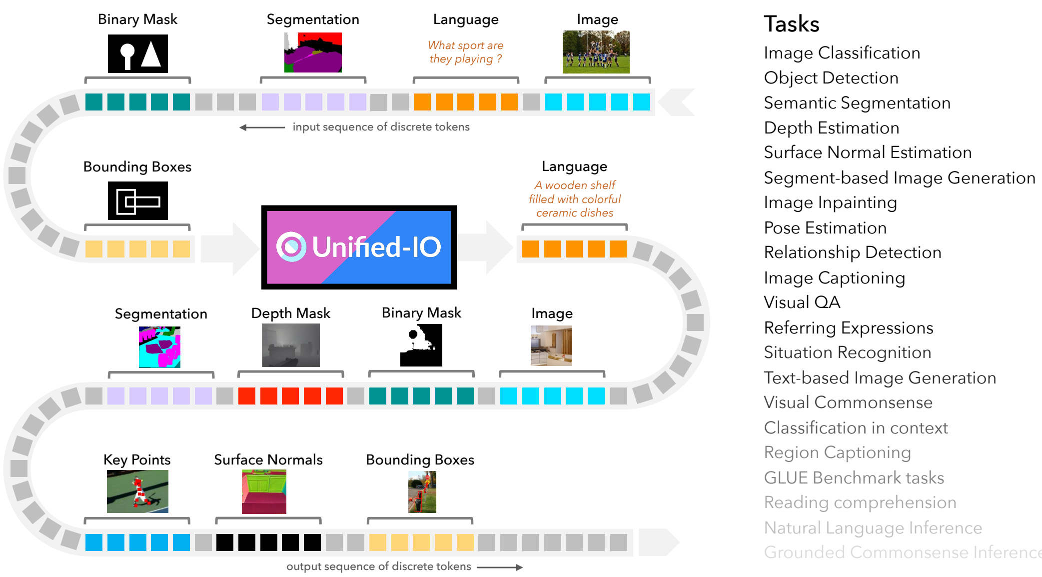


Figure 1: UNIFIED-IO is a sequence-to-sequence model that performs a variety of tasks in computer vision and NLP using a unified architecture without a need for either task or modality specific branches. This broad unification is achieved by homogenizing every task's input and output into a sequence 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.

ABSTRACT

We propose UNIFIED-IO, a model that performs a large variety of AI tasks spanning classical computer vision tasks, including pose estimation, object detection, depth estimation and image generation, vision-and-language tasks such as region captioning and referring expression comprehension, to natural language processing tasks such as question answering and paraphrasing. Developing a single unified model for such a large variety of tasks poses unique challenges due to the heterogeneous inputs and outputs pertaining to each task, including RGB images, per-pixel maps, binary masks, bounding boxes, and language. We achieve this unification by homogenizing every supported input and output into a sequence of discrete vocabulary tokens. This common representation across all tasks allows us to train a single transformer-based architecture, jointly on over 80 diverse datasets in the vision and language fields. UNIFIED-IO is the first model capable of performing all 7 tasks on the GRIT benchmark and produces strong results across 16 diverse benchmarks like NYUv2-Depth, ImageNet, VQA2.0, OK-VQA, Swig, VizWizGround, BoolQ, and SciTail, with no task or benchmark specific fine-tuning. Demos for UNIFIED-IO are available at: unified-io.allenai.org

[*]Equal contribution, corresponding: *jiasenl@allenai.org*

1 INTRODUCTION

We present UNIFIED-IO, the first neural model to jointly perform a large and diverse set of AI tasks spanning classical computer vision (such as object detection, segmentation, and depth estimation), image synthesis (such as image generation and image in-painting), vision-and-language (like visual question answering, image captioning, and referring expression comprehension) and NLP (such as question answering and paraphrasing). Unified general purpose models avoid the need for task specific design, learn and perform a wide range of tasks with a single architecture, can utilize large diverse data corpora, can effectively transfer concept knowledge across tasks and even perform tasks unknown and unobserved at design and training time.

Building unified models for computer vision has proven to be quite challenging since vision tasks have incredibly diverse input and output representations. For instance, object detection produces bounding boxes around objects in an image, segmentation produces binary masks outlining regions in an image, visual question answering produces an answer as text and depth estimation produces a map detailing the distance of each pixel from the camera. This heterogeneity makes it very challenging to architect a single model for all these tasks. In contrast, while the landscape of natural language processing (NLP) tasks, datasets and benchmarks is large and diverse, their inputs and desired outputs can often be uniformly represented as sequences of language tokens. Sequence to sequence (Seq2Seq) architectures (Raffel et al., 2020; Brown et al., 2020), specifically designed to accept and produce such sequences of tokens are thus widely applicable to many tasks. Unified models employing such architectures have been central to much recent progress in NLP.

Unified models for computer vision typically use a shared visual backbone to produce visual embeddings, but then employ individual branches for each of the desired tasks. These include models like Mask-RCNN (He et al., 2017) for classical visual tasks that use an ImageNet pre-trained encoder followed by branches for detection and segmentation, trained in a fully supervised manner. In the vision and language (V&L) domain, CNN backbones feed visual features to transformer architectures that also combine language, followed by task specific heads for visual question answering, referring expression comprehension, visual commonsense reasoning, etc. (Lu et al., 2019; Li et al., 2019; Tan & Bansal, 2019). A more recent trend has seen the emergence of unified architectures that do away with task specific heads and instead introduce modality specific heads (Gupta et al., 2022a; Cho et al., 2021; Kamath et al., 2022) – for instance, a single language decoder that serves multiple tasks requiring language output like captioning and classification. However, most progress in unified models continues to be centered around V&L tasks, owing to the simplicity of building shared language decoders, and is often limited to models that can support just a handful of tasks.

UNIFIED-IO is a Seq2Seq model capable of performing a variety of tasks using a unified architecture without a need for either task or even modality specific branches. This broad unification is achieved by homogenizing every task's output into a sequence of discrete tokens. Dense structured outputs such as images, segmentation masks and depth maps are converted to sequences using a vector quantization variational auto-encoder (VQ-VAE) (Esser et al., 2021), sparse structured outputs such as bounding boxes and human joint locations are transcribed into sequences of coordinate tokens, and language outputs are converted to sequences using byte-pair encoding. This unification enables Unified-IO to jointly train on over 80 datasets spanning computer vision, V&L and NLP tasks with a single streamlined transformer encoder-decoder architecture (Raffel et al., 2020).

Our jointly trained UNIFIED-IO is the first model to support all 7 tasks in the recently announced General Robust Image Task (GRIT) Benchmark (Gupta et al., 2022b). It obtains the state-of-the-art score of 64.26, an accuracy metric averaged across all 7 tasks, handily beating the next best performer, GPV-2 (Kamath et al., 2022), which obtains 32.00. It also easily outperforms GPV-2 on all 4 tasks that GPV-2 can evaluate on. We further evaluate UNIFIED-IO on 16 diverse benchmarks across computer vision and NLP, without any fine-tuning towards any individual benchmark, and find that it performs remarkably well, when compared to specialized (or benchmark fine-tuned) state-of-the-art models.

2 RELATED WORK IN MODEL UNIFICATION

Constructing models that can learn to solve many different tasks has been of long-standing interest to researchers. A traditional approach to this problem is to build models with task-specialized heads on top of shared backbones (Lu et al., 2020a; He et al., 2017). However, this requires manually designing a specialized head for each task, requires one to add new parameters when introducing a new task and potentially limits the ability of the model to transfer between tasks due to the lack of comprehensive parameter sharing.

An alternative is to build *unified* models – models that can complete many different tasks without task-specialized components. In natural language processing this approach has achieved a great deal of success through the use of pre-trained generative models (Raffel et al., 2020; Brown et al., 2020; Chowdhery et al., 2022; Zhang et al., 2022; Rae et al., 2021).

Inspired by this success, there has been a recent trend to build unified models that can be additionally applied to tasks with visual or structured inputs and outputs. A popular design (which we also emulate) has been to encode the inputs and outputs into sequences and then use transformer-based models to parse and make predictions on these sequences. However, models vary greatly in terms of the kinds of inputs and outputs they handle and the details of how they are trained and designed.

Many such models focus on tasks with text and/or image input and text output (Cho et al., 2021; Wang et al., 2022b; Li et al., 2022a). These models use task-specific prompts to signal to the model what task to perform, and can generate arbitrary text as output, allowing them to handle tasks like visual question answering, image classification, and image captioning without task-specific heads. However, only being able to produce natural language output means these models cannot be applied to tasks with structured or visual outputs.

More recent unified models can additionally support image location inputs/outputs, which expands the ranges of tasks those models can perform considerably (e.g., object detection or region captioning). GPV-1 (Gupta et al., 2022a) can generate sets of bounding boxes as output through the use of an end-to-end object detector visual backbone (Carion et al., 2020), and GPV-2 (Kamath et al., 2022) re-uses their text decoder to select bounding boxes to return, and additionally support bounding boxes as input. Another approach is to encode the image positions within the text through the use of special tokens that refer to locations within an image (Cho et al., 2021; Wang et al., 2022a; Chen et al., 2022b). This method is potentially more flexible since it supports points as well as boxes and allows image locations to be interleaved with text. UNIFIED-IO follows this design to encode image positions, but extends it to more tasks including key-point estimation and various tasks that use bounding boxes as inputs (e.g., region classification).

Some recent unified models have extended these capabilities in other directions. Gato (Reed et al., 2022) supports additional modalities including button presses in Atari games or joint torques for a robot arm, and Flamingo (Alayrac et al., 2022) supports interleaved sequences of text, images, and videos as input. However neither of these models support image outputs or image location references limiting the computer vision tasks they can support.

Additional unified models include Perceiver-IO (Jaegle et al., 2021), which supports a range of modalities and proposes a non-auto-regressive decoding approach using task-specific latent query vectors. While effective for some tasks, this method has not been shown to be as effective as auto-regressive decoding on classic generative tasks like captioning or image generation. Uni-Perceiver (Zhu et al., 2022) also supports images, text, and videos and shows good zero-shot performance, but does not support generative tasks.

Concurrent to our work, One-for-All (OFA) (Wang et al., 2022a) proposes a similar approach to UNIFIED-IO that also supports image location tokens and uses a D-VAE to produce images as output. We, however, apply our approach to more tasks (including keypoint estimation, image generation from segmentation, and object detection), and in particularly show how this general approach can be leveraged to handle pixel-labeling vision tasks such as depth estimation, segmentation, and surface normal estimation. Additionally our model differs in several significant respects: We do not use a CNN backbone but instead use a pure transformer approach, we scale our model to a larger size (2.8 billion parameters instead of 930 million), and our pre-training uses de-noising objectives instead of more direct supervision. Importantly, we focus on multi-tasking while OFA is fine-tuned for specific tasks.

Another concurrent work, UViM (Kolesnikov et al., 2022) proposes a unified model for producing visual outputs and applying it to the pixel-labeling tasks of panoptic segmentation, depth prediction, and colorization. Similar to UNIFIED-IO, UViM uses a generative head to predict output tokens that

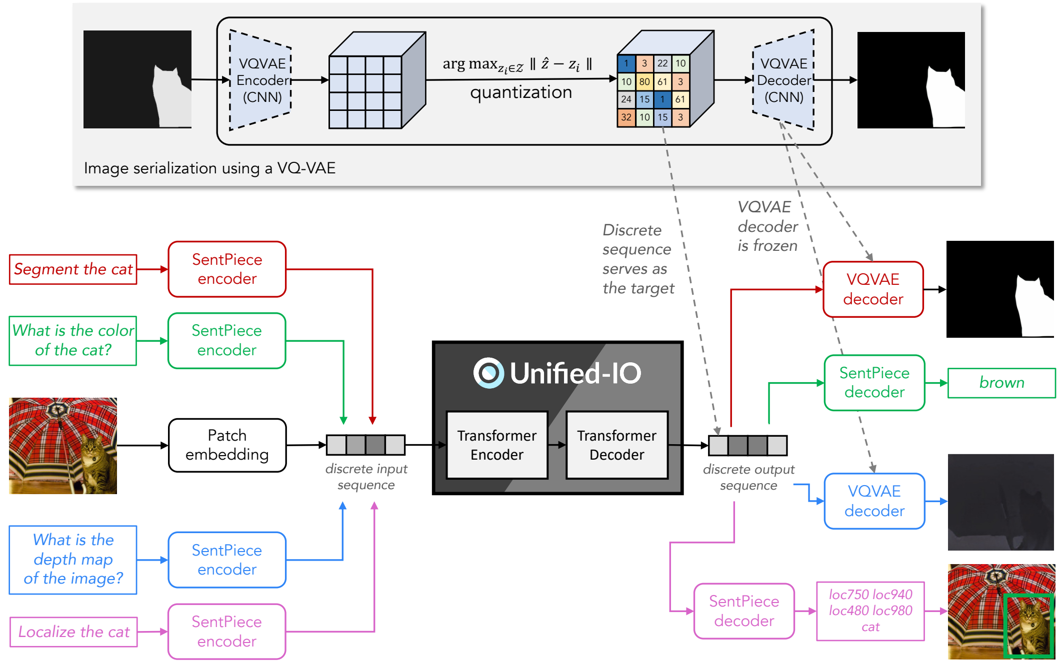


Figure 2: **Unified-IO.** A schematic of the model with four demonstrative tasks: object segmentation, visual question answering, depth estimation and object localization.

are then used as input to a second model to construct an output image. However, UViM learns the second model instead of using a pre-trained D-VAE. Our work additionally covers pure language tasks, language-and-vision tasks, and tasks that require sparse structured output.

3 Unified-IO

3.1 Goal

A unified neural architecture can enable practitioners to train models for new tasks with little to no knowledge of the underlying machinery. Such unified architectures can enable general pre-training that can be beneficial to many downstream applications, reduce the need for task specific parameters, be jointly trained on a large number of tasks and facilitate knowledge transfer from one task to another.

Inputs and outputs across a wide number of tasks in NLP such as question answering, sentiment analysis, paraphrasing and textual entailment can be expressed as sequences of language tokens. Seq2Seq models (Sutskever et al., 2014), built to encode and decode sequences of language tokens, can thus support this large variety of language tasks and can even learn new tasks efficiently without the need to introduce new model parameters.

Computer vision tasks have very diverse input and output structures and do not naturally lend themselves to unification. For example, the task of object detection inputs an image and outputs a set of bounding boxes, each with a label and a score, depth estimation on the other hand, inputs an image and outputs a per-pixel map detailing the distance at that pixel to the camera, and the task of image generation inputs a sequence of text and outputs an RGB image. As a result of such extreme heterogeneity, existing architectures in computer vision tend to rely on highly customized components with tasks specific parameters and designs.

Our goal is to build a single unified and capable model that can support a diverse set of tasks across computer vision and language with little to no need for tasks specific customizations and parameters.

3.2 Unified Task Representations

Supporting a variety of modalities such as images, language, boxes, binary masks, segmentation, etc. without task specific heads, requires us to represent these modalities in a shared and unified space.

We propose discretizing the text, images and other structured outputs in our tasks and representing them with tokens drawn from a unified and finite vocabulary.

Text representation. Each task is specified via a natural language prompt. Prompts are either constructed per task, similar to past works such as McCann et al. (2018); Raffel et al. (2020); Gupta et al. (2022a); Wang et al. (2022a) (e.g., "*What is the depth map of the image?*" for the task of depth estimation) or naturally available as part of the task input (e.g., a natural language question for the task of visual QA). Some tasks also require the model to produce text (e.g., an answer in visual QA or a description in image captioning). Following Raffel et al. (2020), text inputs and outputs are tokenized using SentencePiece (Kudo & Richardson, 2018).

Image representation. A variety of tasks in computer vision require the model to produce high-dimensional outputs such as images. We cast all vision tasks that require dense prediction as a (conditional) image generation problem and unify the output space. Tasks such as image generation and image in-painting require the model to produce an RGB image. Other tasks such as depth estimation and surface normal estimation require the model to produce per pixel maps, which we convert to RGB images. For instance, for depth we construct a gray scale image by normalizing the depth map, while for surface normal estimation we convert the x/y/z orientations into r/g/b values. Finally, tasks such as segmentation require the model to produce a per pixel class label. We convert this format into an RGB image by mapping each class present in the image to a shuffled 8-bit color space. We avoid using a universal color-to-class mapping to prevent the model from being limited to a fixed list of classes, and avoid the image having colors that may only be marginally different due to the presence of a large number of classes.

RGB images can thus support a large collection of tasks in computer vision. These images are encoded using a pre-trained Discrete Variational AutoEncoder (D-VAE) (Van Den Oord et al., 2017). In this work, we use VQ-GAN (Esser et al., 2021) due to its high compression ratio and sharpness. During training we use the pre-trained encoder from VQ-GAN to encode the image into a discrete set of tokens to use as the target sequence. During inference the VQ-GAN decoder is used to convert the sequence of tokens generated by the model into an output image.

Crucially this encoding converts dense 2-d predictions, typically processed and produced by Convolutional Neural Networks (CNNs) into sequences of discrete vocabulary tokens, and can be unified with the encodings produced for language.

Sparse structures representation. Tasks such as object detection require producing bounding boxes around objects whereas tasks like pose estimation require locating a series of joint locations around human beings visible in the image. We take inspiration from recent work that express bounding boxes and image locations as sequences of quantized discrete tokens (Chen et al., 2022b; Clark et al., 2021) and extend this idea to a variety of structured inputs and outputs.

Image locations are encoded with special location tokens. We construct 1000 special tokens which represent discretized coordinates within an image. Points are encoded by using these tokens to encode the x and y coordinates separately (e.g., $\langle \texttt{loc\_500}, \texttt{loc\_500} \rangle$ encodes a point in the middle of the image) and boxes are encoded by specifying the upper right and lower left corners (e.g., $\langle \texttt{loc\_0}, \texttt{loc\_0}, \texttt{loc\_500}, \texttt{loc\_1000} \rangle$ encodes a box over the left half of an image).

3.3 Unified Architecture

Universally representing a wide variety of tasks as input and output sequences of discrete tokens enables us to employ architectures that have been proven successful in natural language processing. In Unified-IO, we propose a pure transformer model largely following the design of T5 (Raffel et al., 2020). In particular, Unified-IO is an encoder-decoder architecture where both the encoder and decoder are composed of stacked transformer layers, which in turn are composed of self-attention transformers, cross-attention transformers (in the decoder), and feed-forward neural networks. The layers are applied residually and layer norms are applied before each transformer and feed-forward network, see Raffel et al. (2020) for details.

We make a few architectural changes to adapt the T5 architecture to our setting. First, to handle 2-d images, we reshape the image into a sequence of flattened 2-d patches and embed it with a linear projection similar to Dosovitskiy et al. (2021). Second, we expand the vocabulary to include the location tokens and image tokens used in the VQ-GAN. Third, we extend the 1-d relative embedding

(Dosovitskiy et al., 2021) to 2-d with a fixed number of learned embeddings. We also add absolute position embedding to the token embedding following Devlin et al. (2019), since the absolute position information is important to image tasks.

In this work, we present four versions of UNIFIED-IO following the corresponding T5 model designs – an `XL` model with 24 layers and 2.8 billion parameters, a `Large` model with 24 layers, 776 million parameters, a `Base` model with 12 layers and 241 million parameters and a `Small` model with 6 layers and 71 million parameters.

3.4 TRAINING

UNIFIED-IO is trained in two stages – A pre-training stage that uses unsupervised losses from text, image and paired image-text data, and a massive multi-task stage where the model is jointly trained on a large variety of tasks. Since our goal is to examine whether a single unified model can solve a variety of tasks simultaneously, we **do not perform task-specific fine-tuning** although prior work (Lu et al., 2020b; Raffel et al., 2020; Wang et al., 2022a) shows it can further improve task performance.

Pre-training. To learn good representations from large-scale webly supervised image and text data, we consider two pre-training tasks: *text span denoising* and *masked image denoising*. The text span denoising task follows Raffel et al. (2020) – randomly corrupt 15% of the tokens and replace the consecutive corrupted tokens with a unique mask token. The masked image denoising task follows Bao et al. (2022) and He et al. (2022) – randomly masked 75% of the image patches, and the goal is to recover the whole images.

We construct the pre-training dataset by incorporating publicly available Language data (i.e., plain texts from Common Crawl[1]), Vision data (i.e., raw images from different datasets) and V&L data (i.e., image caption and image label pairs). For V&L data, we add a simple prompt "*An image of*" at the beginning of caption or categories to indicate it is multi-modal data (Wang et al., 2022b).

Multi-tasking. To build a single unified model for diverse vision, language and V&L tasks, we construct a massive multi-tasking dataset by ensembling over 80 public available datasets. The dataset consists of **classical computer vision tasks** – image classification, object detection, instance segmentation, depth/surface normal estimation; **image synthesis tasks** – image in-painting, image synthesis from caption/segmentation; **V&L Tasks** – visual question answering, image captioning, visual commonsense reasoning and **natural language processing tasks** – GLUE, Squad, etc.

We jointly train UNIFIED-IO on this large set of datasets by mixing examples from these datasets within each batch. For each task category, we sample tasks with a temperature-scaled mixing strategy (Raffel et al., 2020) to make sure the model is sufficiently exposed to underrepresented tasks. We raise each task's mixing rate to the power of $1/T$ and then renormalize the rates so that they sum to 1.

4 TASKS

UNIFIED-IO is jointly trained on a large and diverse set of tasks. In this section, we present many of these tasks with details on how we handle them and provide qualitative examples of both the ground truth and the predictions made by UNIFIED-IO.

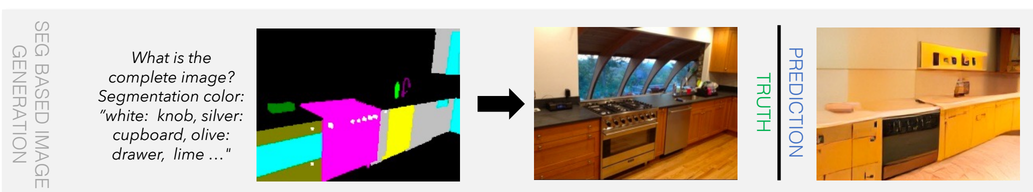
4.1 IMAGE SYNTHESIS TASKS

Image Synthesis from Text. This task requires generating an image that matches a sentence. Training data comes from 4 captioning datasets: COCO Caption (Chen et al., 2015), Conceptual Captions 3M and 12M (Changpinyo et al., 2021), and RedCaps (Desai et al., 2021) as well the datasets used for image classification using the object class as the input caption. Specialized image generation models like DALL·E 2 (Ramesh et al., 2022) use an order of magnitude more data, but we limit our sources to these sets for training efficiency.

[1] http://commoncrawl.org/

Image Inpainting. This task requires filling in a region of an image with a target object. Training data for this task is built from object bounding box annotations from Open Images Kuznetsova et al. (2020), Visual Genome (Krishna et al., 2017) and COCO (Lin et al., 2014). For each object, the input image becomes the source image with the object's bounding box blanked out. The input prompt provides the bounding box's location and the target category. The target output is the original image.

Image Synthesis from Segmentation. This task involves generating an image that matches an input semantic segmentation, i.e., a set of class labels for some or all of the pixels in the image. UNIFIED-IO is trained for this task using segmentation annotations from COCO (Lin et al., 2014), Open Images(Kuznetsova et al., 2020), and LVIS (Gupta et al., 2019) as input. Following the method from Section 3.2 the segmentation input is converted into a RGB image paired with a prompt listing the color-to-class mapping, and the target output is the source image.

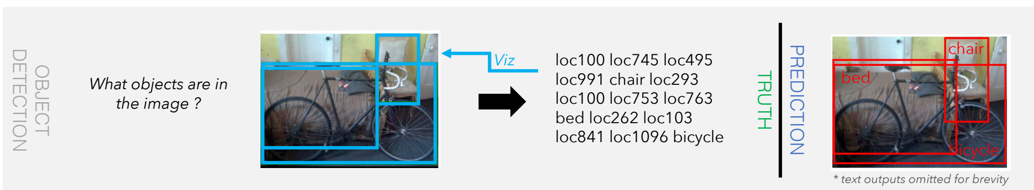


4.2 CLASSICAL COMPUTER VISION TASKS

Image Classification. UNIFIED-IO is trained on 6 image classification datasets: ImageNet 2012 (Deng et al., 2009), ImageNet21k (Ridnik et al., 2021), Places365 (Zhou et al., 2017), Sun397 (Xiao et al., 2010), iNaturalist (Van Horn et al., 2018) and Caltech bird (Welinder et al., 2010). For this task the input is an image and a static prompt, and the output is a class name. During inference we compute the log-probability of each class label in the dataset being evaluated and return the highest scoring one. This ensures UNIFIED-IO does not return a category from a different categorization dataset that is a synonym or hypernym of the correct label.

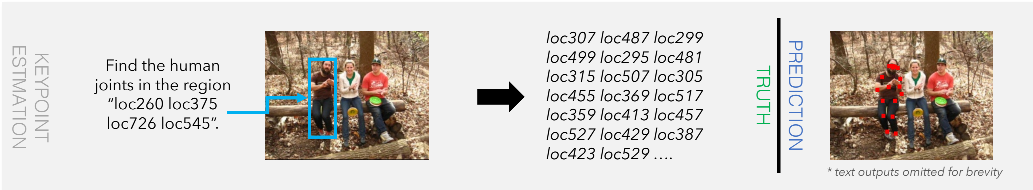
Object Categorization. This task identifies which label, from a given set, best corresponds to an image region defined by an input image and bounding box. The input is the image, a prompt specifying the image region and the output is the target class name. We convert object detection annotations from Visual Genome, Open Images, and COCO for this task. Inference is constrained to return a valid label for the target label set just as with image classification.

Object Detection. UNIFIED-IO is trained on object detection annotations from Visual Genome (Krishna et al., 2017), Open Images (Kuznetsova et al., 2020), and COCO Lin et al. (2014). For this task the input is a static prompt and an image, and the output text includes the bounding boxes and class names of all objects in the image. We randomize the order of the output objects during training, but for simplicity leave integrating more complex data-augmentation techniques (Chen et al., 2022b) to future work.

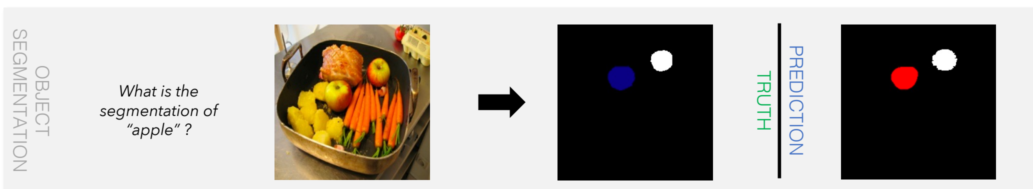


Object Localization. Object localization requires returning bounding boxes around all objects of a given category. Training data is derived from our object detection training data by constructing a training example from each category of objects present in an image. The input is then the image, a prompt specifying the target class, and the output is a list of all boxes that contain an instance of that class. The class for each box (which is always the class specified in the prompt) is included in the output for the sake of keeping the output format consistent with the object detection output. Object localization can use input categories which are not present in the image. To handle this, we construct negative samples by randomly selecting categories not present in the image to use as input, in which case the output is an empty sequence.

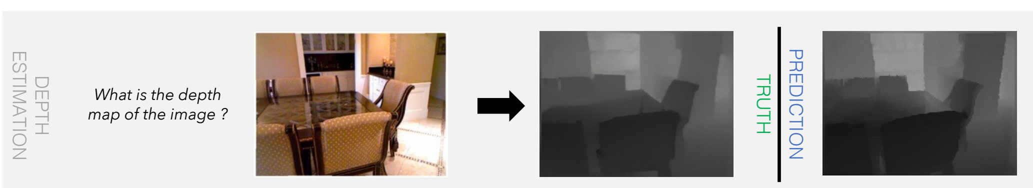
Keypoint Estimation. Keypoint estimation requires returning the location of 17 keypoints on a human body (e.g., eyes, nose, feet, etc.) for each person in an image. While it is possible to perform this task in one pass by listing the keypoints of all people in the image in a single output sequence, this can result in an extremely long output sequence, so UNIFIED-IO uses a multi-step approach for this task. To do this UNIFIED-IO is trained to complete the subtask of detecting the keypoints for single a person in a given region. For this subtask, the input prompt specifies the target region and and the output is a list of 17 points (a pair of locations tokens for the x and y coordinates) along with a visibility labels (1 for not visible, 2 for partly visible, 2 for fully visible). Non-visible are always given the coordinates (0, 0), and during inference those points are replaced with the mean of the visible points to provide as best-guess in case that point is marked as visible in the ground truth. Training data for this subtask comes from COCO human pose data (Lin et al., 2014) with the ground-truth person regions as input. During inference we locate person regions using the object localization prompt, then apply UNIFIED-IO again to find keypoints for each detected region.



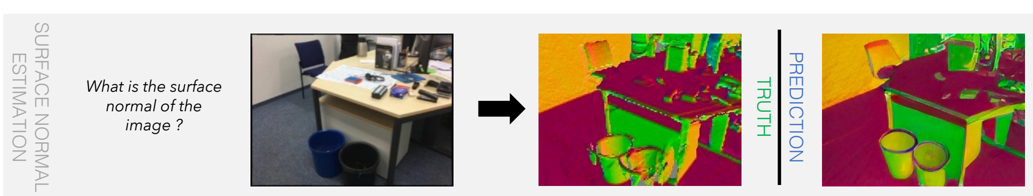
Object Segmentation. Object segmentation requires finding the binary segmentation mask of each instance of a particular category in an image. The input is an image and a prompt that includes the target class, while the output is an RGB image with black background and instances of that class filled in with unique colors following the method in Section 3.2. The output image is resized to match the input image if needed using a nearest-neighbor resizing method, and binary masks are built from each unique color. In practice the output image from UNIFIED-IO can have slightly non-uniform colors or extraneous background pixels, likely due to limitation in what the D-VAE can decode/encode, so the output pixels are clustered by color and and connected components of less than 8 pixels are removed to build cleaned instance masks. Segmentation annotations come from Open Images (Kuznetsova et al., 2020), LVIS (Gupta et al., 2019), and COCO (Lin et al., 2014).



Depth Estimation. Depth estimation requires assigning each pixel in an image a depth value. This task uses a static prompt as input, and the output is a grayscale image representing the normalized depth at each pixel. The generated output image is reiszed to the same size as the input image and then pixel values are rescaled to the maximum depth in the training to get an output depth map. Training data comes from the NYU Depth Dataset V2 (Nathan Silberman & Fergus, 2012).



Surface Normal Estimation. UNIFIED-IO is trained on FrameNet (Huang et al., 2019a) and BlendedMVS (Yao et al., 2020) surface normal estimation datasets. For this task the input is a static prompt and an image, the output is an RGB representation of the x/y/z orientation of the surface at each pixel. The generated output image is resized to match the input image and converted back to x/y/z orientations to produce the final output.



4.3 VISION & LANGUAGE TASKS

Visual Question Answering. UNIFIED-IO is trained on a collection of VQA datasets including VQA 2.0 (Goyal et al.), Visual Genome, VizWizVQA (Gurari et al., 2018) and A-OKVQA (Schwenk et al., 2022). For VQA, the question is used as the prompt, and the output is the answer text. For VQA, it is common to constrain the model to predict an answer from a fixed last of common VQA answers (Wang et al., 2022a;b) during inference, but we avoid doing this since we find it does not benefit UNIFIED-IO in practice.

Answer-Grounded Visual Question Answering. This task requires both answering a question and returning a binary mask specifying the region of the image used to answer the question. The format for this task follows the one for VQA except that a binary mask is also used as an additional output. Training data comes from VizWiz-VQA (Chen et al., 2022a), a dataset designed to train models that could benefit people with visual impairments.

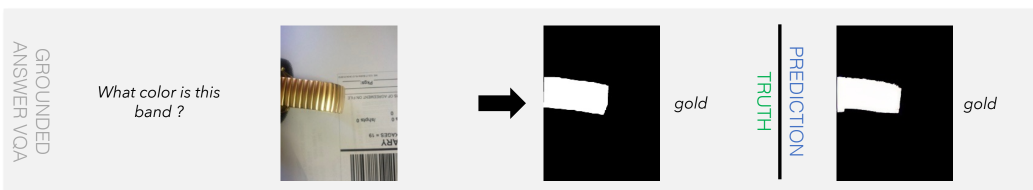
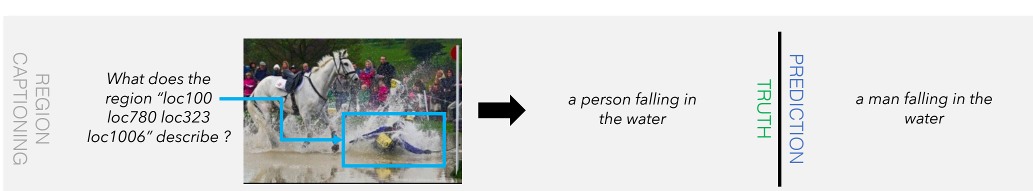


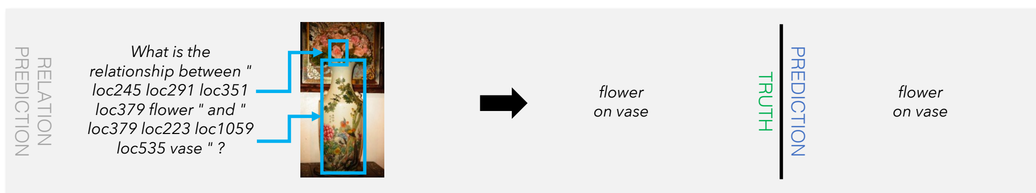
Image Captioning. Image captioning data comes from the same manually annotated and unsupervised sources used for Image Generation. In this case the inputs and output are reversed, the input is an image and the static prompt, and the output is a caption that matches the image.

Region Captioning. Region captioning tasks a model with generating a caption that describes a specific region in the image. Our format for this task is identical to Image Captioning except the region is included in the input prompt. Visual Genome (Krishna et al., 2017) is used for the training data.



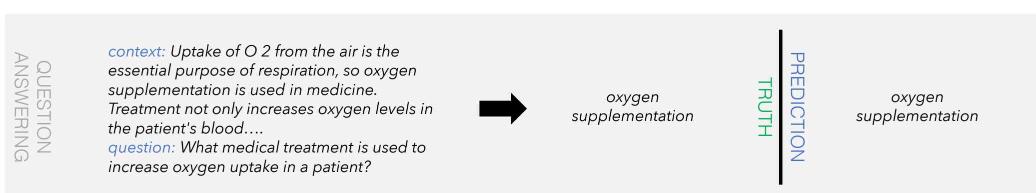
Referring Expression Comprehension. The task requires the model to localize an image region described by a natural language expression. The annotation is similar to Object Localization, except that the target is specified with natural language expression instead of class name. Datasets for this task include RefCOCO (Kazemzadeh et al., 2014), RefCOCO+ (Kazemzadeh et al., 2014) and RefCOCOg (Mao et al., 2016).

Relationship Detection. This task requires predicating a relationship between a pair of objects which are grounded by bounding boxes. The prompt contains both the object regions, and the output is the predicted predicate. There are 2 datasets in this tasks: Visual Genome (Krishna et al., 2017) and Open Images (Kuznetsova et al., 2020).

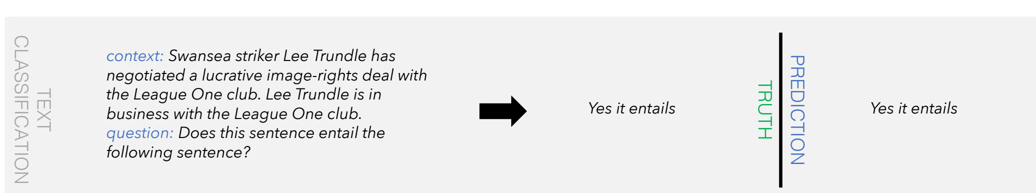


4.4 Natural Language Processing Tasks

Question Answering. Following prior work in natural language processing (Raffel et al., 2020), QA tasks are formatted by placing both the question and any text context (e.g., an paragraph containing the answer) into the prompt and training the model to generate the text answer. Unified-IO is trained on several QA datasets including SQuAD (Rajpurkar et al., 2016), SWAG (Zellers et al., 2018), OpenBookQA (Mihaylov et al., 2018), Common Sense QA (Talmor et al., 2018), Freebase QA (Jiang et al., 2019), Fever 1.0 (Thorne et al., 2018), SciQ (Welbl et al., 2017) and CosmoQA (Huang et al., 2019b).



Text Classification. Also following past work (Raffel et al., 2020), text classification tasks are formatted by placing the input sentences and a query in the prompt and training the model to generate the target class. Datasets include tasks from GLUE (Wang et al., 2018), SuperGLUE (Wang et al., 2019), SNLI (Bowman et al., 2015) and SciTail (Khot et al., 2018).



		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 [3]	-	-	-	-	-	-	-	-	-	-	-	-	**49.6**	**50.5**	7.2	7.1
1	Mask R-CNN [29]	-	-	44.7	45.1	-	-	-	-	26.2	26.2	**70.8**	**70.6**	-	-	20.2	20.3
2	GPV-1 [26]	33.2	33.2	42.8	42.7	50.6	49.8	25.8	26.8	-	-	-	-	-	-	21.8	21.8
3	CLIP [56]	48.1	-	-	-	-	-	-	-	-	-	-	-	-	-	6.9	-
4	OFA$_{\text{LARGE}}$ [73]	22.6	-	-	-	72.4	-	61.7	-	-	-	-	-	-	-	22.4	-
5	GPV-2 [36]	54.7	55.1	53.6	53.6	63.5	63.2	51.5	52.1	-	-	-	-	-	-	31.9	32.0
6	UNIFIED-IO$_{\text{SMALL}}$	42.6	-	50.4	-	52.9	-	51.1	-	40.7	-	46.5	-	33.5	-	45.4	-
7	UNIFIED-IO$_{\text{BASE}}$	53.1	-	59.7	-	63.0	-	68.3	-	49.3	-	60.2	-	37.5	-	55.9	-
8	UNIFIED-IO$_{\text{LARGE}}$	57.0	-	64.2	-	67.4	-	74.1	-	54.0	-	67.6	-	40.2	-	57.0	-
9	UNIFIED-IO$_{\text{XL}}$	**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**

Table 1: Comparison of our UNIFIED-IO models to recent SOTA on GRIT benchmark. UNIFIED-IO is the first model to support all seven tasks in GRIT.

5 EXPERIMENTS

We are interested in building a unified model that can perform a wide variety of tasks by simply prompting it for the desired task. Hence, we jointly train UNIFIED-IO on over 80 vision and NLP datasets and **do not fine tune** it for any specific task or benchmark. We now present results for UNIFIED-IO on the GRIT test benchmark (Sec 5.1), ablate model sizes and training data via the GRIT ablation benchmark (Sec 5.2) and evaluate UNIFIED-IO on 16 other benchmarks in computer vision and NLP (Sec 5.3). Qualitative results can be found in Section 3.2 as well as our demo[2].

5.1 RESULTS ON GRIT

The General Robust Image Task (GRIT) Benchmark (Gupta et al., 2022b) is an evaluation-only benchmark designed to measure the performance of models across multiple tasks, concepts, and data sources. GRIT aims to encourage the building of unified and general purpose vision models, and is thus well suited to evaluate UNIFIED-IO. GRIT also makes a large range of metrics available that measure the performance, robustness, and calibration of a vision system. GRIT has seven tasks that cover a range of visual skills with varying input and output modalities and formats: *object categorization*, *object localization*, *referring expression grounding*, *visual question answering*, *segmentation*, *human keypoint detection*, and *surface normal estimation*. See Section 3.2 for qualitative examples and details on how UNIFIED-IO handles these task. Note that UNIFIED-IO must be evaluated on GRIT's *Unrestricted* track due to the datasets used for pre-training and multi-task training. Also, the quantitative results presented here are obtained by our multi-task trained model without any further finetuning on GRIT related tasks or data sources.

UNIFIED-IO is the first model to support all seven tasks in GRIT. As seen in Table 1, UNIFIED-IO$_{\text{XL}}$ outperforms all prior submissions to GRIT obtaining 64.3 on test, the average accuracy across all 7 tasks. The next best submission is GPV-2 (Kamath et al., 2022) that obtains 32.0 and can only support 4 out of 7 tasks. Importantly, UNIFIED-IO$_{\text{XL}}$ also beats GPV-2 on each of these 4 tasks, by large margins of over 5 points on categorization, over 13 points on localization, 11 points on VQA and 26 points on referring expressions.

Another recent unified model evaluated on GRIT is OFA (Wang et al., 2022a) which supports 3 tasks. GRIT is able to comfortably outperform the multi-task checkpoint of OFA on these 3 tasks. Mask R-CNN can perform 3 tasks as well (localization, segmentation and keypoints). Interestingly, UNIFIED-IO$_{\text{XL}}$ outperforms Mask-RCNN on 2 tasks and gets to within 3 points on the keypoint estimation task, even though Mask-RCNN uses task specific network heads for those 3 tasks, where as UNIFIED-IO has no such specialized parameters.

NLL-AngMF (Bae et al., 2021) is a SOTA model for surface normal estimation but can only perform that 1 task on GRIT. On surface normal estimation, UNIFIED-IO gets 44.3 vs. 49.6 for NLL-AngMF, a reasonable result given our model is pure transformer model without a pyramid network or custom loss functions. A limitation of our method for this task is that our VQVAE cannot construct surface

[2] unified-io.allenai.org

		restricted	params (M)	Categorization		Localization		VQA		Refexp		Segmentation		Keypoint		Normal	
				same	new	same	new	same	new	same	new	same	new	same	new	same	new
0	NLL-AngMF	✓	72	-	-	-	-	-	-	-	-	-	-	-	-	**50.7**	-
1	Mask R-CNN	✓	58	-	-	51.9	40.8	-	-	-	-	44.9	0.3	**70.9**	-	-	-
2	GPV-1	✓	236	58.7	0.8	48.3	37.8	58.4	74.0	29.7	23.1	-	-	-	-	-	-
3	CLIP		302	49.1	46.7	-	-	-	-	-	-	-	-	-	-	-	-
4	$\text{OFA}_{\text{LARGE}}$		473	28.9	15.8	-	-	74.9	88.6	63.4	58.5	-	-	-	-	-	-
5	GPV-2		370	**85.0**	13.5	54.6	54.2	69.8	81.7	57.8	48.3	-	-	-	-	-	-
6	$\text{UNIFIED-IO}_{\text{SMALL}}$		71	52.9	31.9	47.5	61.5	59.0	72.5	54.2	45.7	37.4	48.5	46.6	-	33.6	-
7	$\text{UNIFIED-IO}_{\text{BASE}}$		241	60.3	47.5	57.9	68.4	68.0	81.8	72.5	62.2	45.8	57.2	60.2	-	37.7	-
8	$\text{UNIFIED-IO}_{\text{LARGE}}$		776	63.0	52.7	63.3	70.9	72.1	84.3	79.2	66.3	50.4	62.2	67.7	-	40.3	-
9	$\text{UNIFIED-IO}_{\text{XL}}$		2925	66.1	**60.1**	**65.6**	**74.4**	**78.6**	**90.2**	**83.5**	**72.4**	**53.0**	**64.2**	68.2	-	45.1	-

Table 2: Generalization to new concepts on the GRIT ablation set.

normal images with perfect precision. For example, on Framenet (Kazemzadeh et al., 2014) we find the upper bound of what our method can achieve using the VQVAE is 59.8. This suggest our score could be considerably improved by training a stronger VQVAE for these kinds of images.

Evaluations on *same* concept and *new* concepts. GRIT provides a breakdown of metrics into two groups: *same* for samples that only contain concepts seen in the primary training data (a set of common datasets like COCO, ImageNet and Visual Genome), and *new* for samples containing at least one concept unseen in primary training data. Table 2 shows results for UNIFIED-IO and other leaderboard entries for the ablation set, divided into same and new concepts.

$\text{UNIFIED-IO}_{\text{XL}}$ shows little degradation in performance between *same* and *new*, compared to competing entries. On some tasks UNIFIED-IO is even able to outperform on the *new* split compared to the *same*. This indicates that the volume of training data used to train UNIFIED-IO has a broad coverage of concepts, and provides almost as effective a level of supervision as provided by large standard vision datasets like COCO. Furthermore, since UNIFIED-IO is a uniquely unified architecture with no task specific parameters, it is very likely able to effectively transfer knowledge across different tasks.

In comparison to Mask-RCNN (row 1), GRIT metrics show UNIFIED-IO (row 14) is better by a large margin on *new* concepts, i.e., non-COCO examples (74.4 vs 40.8 for localization and 64.2 vs 0.3 on segmentation), but is still superior on the COCO-like examples (65.6 vs 51.9 for localization and 53.0 vs 44.9 on segmentation). UNIFIED-IO is also able to beat GPV-2 (row 5) on *new* concepts by large margins across all 4 tasks supported by GPV-2 even though GPV-2 is exposed to these concepts via webly supervised data and is designed to transfer concept knowledge across skills.

5.2 ABLATIONS

Table 1 compares the 4 size variants of UNIFIED-IO: `Small`, `Base`, `Large` and `XL`, with 71M, 241M, 776M and 2.8B parameters respectively. We see considerable gains when moving from the 71M parameter `Small` to the 2.8B parameter `XL` variant. There is no sign that performance is saturating, suggesting UNIFIED-IO's performance would further improve with an even larger model. We leave additional ablations to future work.

Table 2 compares these size variants for the *same* and *new* concepts in GRIT. Here we notice that the gaps between *same* and *new* shrink as we move to larger models. For example, `Small` (row 6) shows a difference of 21 pts on Categorization while the `XL` model's difference (row 9) is 6. This indicates that the larger models are able to learn a large number of concepts well, and the performance gap between more common and less common concepts begins to shrink as the model scales.

5.3 RESULTS ON ADDITIONAL TASKS

We report results on 16 additional tasks used in our multi-task training setup. For these tasks, we do not expect to get state-of-the-art results since specialized models are usually designed and hyper-parameter tuned for a single task, while we are evaluating a single jointly trained model. We also

	NYUv2	*ImageNet*	*Place365*	*VQAv2*	*OkVQA*	*A-OkVQA*	*VizWizQA*	*VizWizGround*	*Swig*	*SNLI-VE*	*VisComet*	*Nocaps*	*COCO*	*COCO*	*MRPC*	*BoolQ*	*SciTail*
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	-	-	-	Flamingo	-	Flamingo	-	-	-	-	-	-	-	T5	PaLM	-
	0.467	-	-	-	57.8	-	49.8	-	-	-	-	-	-	-	92.20	92.2	-
UNIFIED-IO$_{\text{SMALL}}$	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$_{\text{BASE}}$	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$_{\text{LARGE}}$	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$_{\text{XL}}$	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	CoCa	MAE	CoCa	KAT	GPV2	Flamingo	MAC-Caps	JSL	OFA	SVT	CoCa	-	OFA	Turing NLR	ST-MOE	DeBERTa
	0.330	91.00	60.3	82.3	54.4	38.1	65.7	27.3	39.6	91.0	18.3	122.4	-	145.3	93.8	92.4	97.7

Table 3: Comparing the jointly trained UNIFIED-IO to specialized and benchmark fine-tuned state of the art models across Vision, V&L and Language tasks. Benchmarks used for evaluation are: NYUv2 (Nathan Silberman & Fergus, 2012), ImageNet (Deng et al., 2009), Places365 (Zhou et al., 2017), VQA 2.0 (Goyal et al.), A-OKVQA (Schwenk et al., 2022), VizWizVQA (Gurari et al., 2018), VizWizGround (Chen et al., 2022a), Swig (Pratt et al., 2020), SNLI-VE (Xie et al., 2019), VisComet (Park et al., 2020), Nocaps (Agrawal et al., 2019), COCO Captions (Chen et al., 2015), MRPC (Dolan & Brockett, 2005), BoolQ (Clark et al., 2019), and SciTail (Khot et al., 2018).

avoid extensive task-specific tricks like color jitter, horizontal flipping, CiDER optimization and label smoothing, which are often responsible for considerable gains on individual task performance. We leave such task specific tuning for future work. See Table 3 for the results. When possible, we additionally report the best prior result on these tasks from a unified model, meaning a model that is trained in a multi-task setting and a unified architecture (no task specific head or customizations) with at least three other tasks.

UNIFIED-IO provides strong performance on all these tasks despite being massively multi-tasked. We review more fine-grained results below.

Depth Estimation. On depth estimation, UNIFIED-IO achieves 0.385 rmse, which is behind state-of-the-art (Li et al., 2022b) but ahead of recently proposed unified model, UViM Kolesnikov et al. (2022), despite being trained to do far more tasks.

Image Classification. UNIFIED-IO achieves 79.1 on Imagenet and 53.2 on Places365, showing the model was able to retain knowledge of many fine-grained classes despite being massively multi-tasked. Notably, we achieve this without the extensive data augmentations methods typically used by state-of-the-art models (Yu et al., 2022; He et al., 2022).

Visual Question Answering. UNIFIED-IO is competitive with fine-tuned models on VQA (Alayrac et al., 2022; Kamath et al., 2022; Gui et al., 2021), and achieves state-of-the-art results on A-OKVQA. Relative to multi-task Flamingo, UNIFIED-IO performs better on VizWiz-QA but worse on OK-VQA.

Image Captioning. Despite the lack of CiDER optimize, UNIFIED-IO is a strong captioning model, and generalizes well to nocaps. Since UNIFIED-IO is trained on many captioning datasets, it is likely the use of style tags following Cornia et al. (2021) would offer additional improvement by signaling UNIFIED-IO to specifically generate COCO-style captions during inference.

NLP tasks.: UNIFIED-IO achieves respectable results on three NLP tasks but lags behind state-of-the-models (Smith et al., 2022; Zoph et al., 2022; He et al., 2020). This can at least partly be attributed to scale, modern NLP models contain 100 billion+ parameters and do much more extensive NLP pre-training.

6 Conclusion

We have presented UNIFIED-IO, a unified architecture that supports a variety of tasks spanning computer vision, NLP and vision-and-language. These tasks require diverse inputs and outputs including images, depth-maps, surface normal maps, binary masks, segmentation masks, text, bounding boxes and keypoints. This unification is made possible by homogenizing each of these modalities into a sequence of discrete tokens. We train a 2.8 billion parameter UNIFIED-IO model jointly on over 80 data sources. This joint model is the first model to perform all 7 tasks on the GRIT benchmark and obtains impressive results across 16 other vision and NLP benchmarks, with no benchmark fine-tuning or task-specific modifications.

ACKNOWLEDGEMENTS

Research supported with Cloud TPUs from Google's TPU Research Cloud (TRC). We thank Zak Stone and the Google Cloud TPU team for providing access to the TPU machines used for conducting experiments and Keisuke Sakaguchi for providing early feedback on the project. We also thank the ReVIZ team at the Allen Institute for AI, including Sam Stuesser, Sam Skjonsberg, Jon Borchardt, Carissa Schoenick and Michael Schmitz for helping setup the demo website: `unified-io.allenai.org`.

REFERENCES

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. nocaps: novel object captioning at scale. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 8948–8957, 2019. 14

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. *arXiv*, 2022. 3, 14

Gwangbin Bae, Ignas Budvytis, and Roberto Cipolla. Estimating and exploiting the aleatoric uncertainty in surface normal estimation. In *ICCV*, 2021. 12

Hangbo Bao, Li Dong, and Furu Wei. BEiT: BERT pre-training of image transformers. In *ICLR*, 2022. 6

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. The SNLI corpus. 2015. 11

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *NeurIPS*, 2020. 2, 3

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 3

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 2021. 6

Chongyan Chen, Samreen Anjum, and Danna Gurari. Grounding answers for visual questions asked by visually impaired people. In *CVPR*, 2022a. 10, 14

Ting Chen, Saurabh Saxena, Lala Li, David J Fleet, and Geoffrey Hinton. Pix2seq: A language modeling framework for object detection. In *ICLR*, 2022b. 3, 5, 8

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv*, 2015. 6, 14

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *ICML*, 2021. 2, 3

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv*, 2022. 3

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. *arXiv preprint arXiv:1905.10044*, 2019. 14

Christopher Clark, Jordi Salvador, Dustin Schwenk, Derrick Bonafilia, Mark Yatskar, Eric Kolve, Alvaro Herrasti, Jonghyun Choi, Sachin Mehta, Sam Skjonsberg, Carissa Schoenick, Aaron Sarnat, Hannaneh Hajishirzi, Aniruddha Kembhavi, Oren Etzioni, and Ali Farhadi. Iconary: A pictionary-based game for testing multimodal communication with drawings and text. In *EMNLP*, 2021. 5

Marcella Cornia, Lorenzo Baraldi, Giuseppe Fiameni, and Rita Cucchiara. Universal captioner: long-tail vision-and-language model training through content-style separation. *arXiv preprint arXiv:2111.12727*, 2021. 14

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009. 7, 14

Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. RedCaps: Web-curated image-text data created by the people, for the people. *arXiv*, 2021. 6

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019. 6

William B. Dolan and Chris Brockett. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*, 2005. URL https://aclanthology.org/I05-5002. 14

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 5, 6

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021. 2, 5

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. 10, 14

Liangke Gui, Borui Wang, Qiuyuan Huang, Alex Hauptmann, Yonatan Bisk, and Jianfeng Gao. Kat: A knowledge augmented transformer for vision-and-language. *arXiv preprint arXiv:2112.08614*, 2021. 14

Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019. 7, 9

Tanmay Gupta, Amita Kamath, Aniruddha Kembhavi, and Derek Hoiem. Towards general purpose vision systems: An end-to-end task-agnostic vision-language architecture. In *CVPR*, 2022a. 2, 3, 5, 12

Tanmay Gupta, Ryan Marten, Aniruddha Kembhavi, and Derek Hoiem. GRIT: General robust image task benchmark. *arXiv*, 2022b. 2, 12

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3608–3617, 2018. 10, 14

Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017. 2, 3, 12

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 6, 14

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. DeBERTa: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*, 2020. 14

Jingwei Huang, Yichao Zhou, Thomas Funkhouser, and Leonidas J Guibas. Framenet: Learning local canonical frames of 3d surfaces from a single rgb image. In *ICCV*, 2019a. 9

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Cosmos qa: Machine reading comprehension with contextual commonsense reasoning. *arXiv*, 2019b. 11

Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, et al. Perceiver io: A general architecture for structured inputs & outputs. *arXiv preprint arXiv:2107.14795*, 2021. 3

Kelvin Jiang, Dekun Wu, and Hui Jiang. Freebaseqa: A new factoid qa data set matching trivia-style question-answer pairs with freebase. In *NAACL-HLT*, 2019. 11

Amita Kamath, Christopher Clark, Tanmay Gupta, Eric Kolve, Derek Hoiem, and Aniruddha Kembhavi. Webly supervised concept expansion for general purpose vision models. *arXiv*, 2022. 2, 3, 12, 14

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, 2014. 11, 13

Tushar Khot, Ashish Sabharwal, and Peter Clark. Scitail: A textual entailment dataset from science question answering. In *AAAI*, 2018. 11, 14

Alexander Kolesnikov, André Susano Pinto, Lucas Beyer, Xiaohua Zhai, Jeremiah Harmsen, and Neil Houlsby. Uvim: A unified modeling approach for vision with learned guiding codes. *arXiv preprint arXiv:2205.10337*, 2022. 3, 14

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2017. 7, 8, 10, 11

Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *EMNLP*, 2018. 5

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 2020. 7, 8, 9, 11

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022a. 3

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. In *arXiv*, 2019. 2

Zhenyu Li, Xuyang Wang, Xianming Liu, and Junjun Jiang. Binsformer: Revisiting adaptive bins for monocular depth estimation. *arXiv preprint arXiv:2204.00987*, 2022b. 14

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 7, 8, 9

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 2019. 2

Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, D. Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10434–10443, 2020a. 3

Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *CVPR*, 2020b. 6

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, 2016. 11

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. The natural language decathlon: Multitask learning as question answering. *arXiv*, 2018. 5

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *EMNLP*, 2018. 11

Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from RGBD images. In *ECCV*, 2012. 9, 14

Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. Visualcomet: Reasoning about the dynamic context of a still image. In *In Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 14

Sarah Pratt, Mark Yatskar, Luca Weihs, Ali Farhadi, and Aniruddha Kembhavi. Grounded situation recognition. In *European Conference on Computer Vision*, pp. 314–332. Springer, 2020. 14

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021. 12

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*, 2021. 3

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 2020. 2, 3, 5, 6, 11

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. In *EMNLP*, 2016. 11

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv*, 2022. 6

Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, Tom Eccles, Jake Bruce, Ali Razavi, Ashley Edwards, Nicolas Heess, Yutian Chen, Raia Hadsell, Oriol Vinyals, Mahyar Bordbar, and Nando de Freitas. A generalist agent. *arXiv*, 2022. 3

Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv*, 2021. 7

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. *arXiv*, 2022. 10, 14

Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, et al. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. *arXiv preprint arXiv:2201.11990*, 2022. 14

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014. 4

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *NAACL*, 2018. 11

Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *EMNLP*, 2019. 2

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. Fever: a large-scale dataset for fact extraction and verification. In *NAACL*, 2018. 11

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *NeurIPS*, 2017. 5

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *CVPR*, 2018. 7

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv*, 2018. 11

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. In *NeurIPS*, 2019. 11

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *arXiv*, 2022a. 3, 5, 6, 10, 12

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. In *ICLR*, 2022b. 3, 6, 10

Johannes Welbl, Nelson F. Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. In *NUT@EMNLP*, 2017. 11

P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010. 7

Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010. 7

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*, 2019. 14

Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *CVPR*, 2020. 9

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 14

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. Swag: A large-scale adversarial dataset for grounded commonsense inference. In *EMNLP*, 2018. 11

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022. 3

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *TPAMI*, 2017. 7, 14

Xizhou Zhu, Jinguo Zhu, Hao Li, Xiaoshi Wu, Hongsheng Li, Xiaohua Wang, and Jifeng Dai. Uni-perceiver: Pre-training unified architecture for generic perception for zero-shot and few-shot tasks. In *CVPR*, 2022. 3

Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam Shazeer, and William Fedus. Designing effective sparse expert models. *arXiv preprint arXiv:2202.08906*, 2022. 14