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

Anonymous authors
Paper under double-blind review

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, 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 90 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-specific fine-tuning. Code and pre-trained models will be made publicly available.

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) 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 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 R-CNN (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,

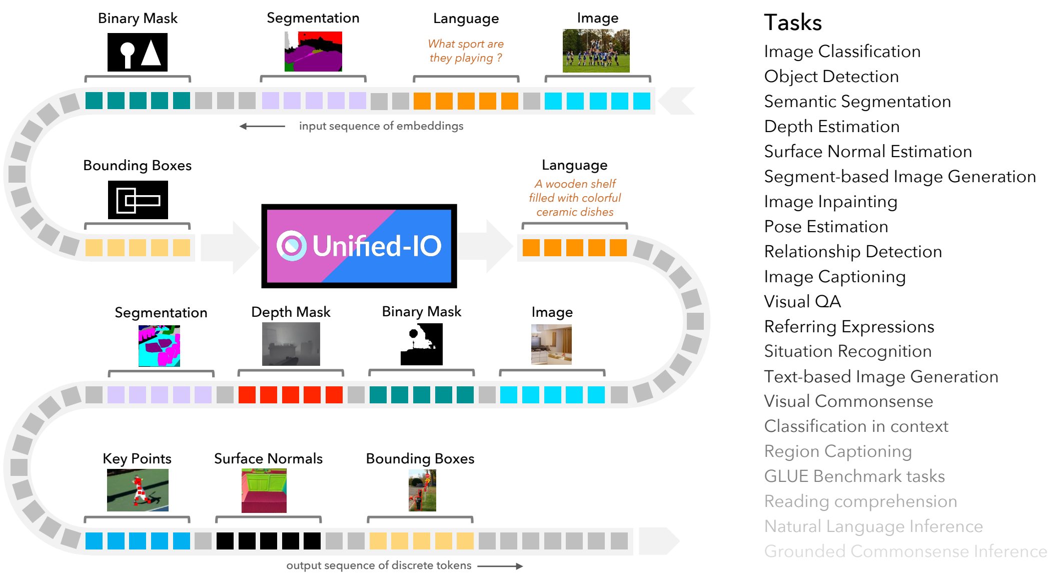

Figure 1: UNIFIED-IO is a single sequence-to-sequence model that performs a variety of tasks in computer vision and NLP using a unified architecture without a need for either task or modality-specific branches. This broad unification is achieved by homogenizing every task's input and output into a sequence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

referring expression, 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 (Hu & Singh, 2021; Cho et al., 2021; Gupta et al., 2022a; Wang et al., 2022b) – 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 supporting 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 90 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 General Robust Image Task (GRIT) Benchmark (Gupta et al., 2022b) and obtains the top overall score of 64.3 when averaging across all tasks, handily beating the second best model by 32.0. 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 compared to specialized (or fine-tuned) state-of-the-art models.

2 VISION, LANGUAGE AND MULTI-MODAL TASKS

UNIFIED-IO is designed to handle a wide range of language, vision and language, and classic vision tasks in a unified way. To fully test this capability, we gather 95 vision, language, and multi-modal datasets from 62 publicly available data sources as targets for our model to learn during multi-task training. These datasets cover a wide range of tasks, skills, and modalities.

	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**	✓	✓	✓	✓	✓	✓	✓	✓

Table 1: Tasks UNIFIED-IO learns to complete. From left to right, columns show an example of one of the sources used for the task, the number of datasets, total number and percent of examples relative to the entire training corpus, and sample rate during multi-task training. Subsequent columns show what modalities are required for the tasks, and highlighted rows show aggregated statistics for groups of similar tasks.

We categorize the input and output modalities of each task into 4 different types: `Text` – natural language tokens; `Image` – RGB images; `Sparse` – a small number of location coordinates within the image; `Dense` – per-pixel labels such as depth maps, surface normal maps, *etc*. We group related datasets into 8 groups and 22 tasks to facilitate our training and analysis:

Image Synthesis. Given a text description, partially occluded image and inpainting target, or segmentation map containing a semantic class for some pixels, generate a matching image. Data sources with image and text pairs (Desai et al., 2021), bounding boxes (Krishna et al., 2017) or semantic segmentation (Gupta et al., 2019) can be used to build these tasks.

Sparse Labelling. Given an image and a natural language query, identify the target regions or keypoint locations that are being referred to. Tasks include object detection (Kuznetsova et al., 2020), object localization (Rhodes et al., 2017), human pose estimation (Lin et al., 2014) and referring expression (Kazemzadeh et al., 2014).

Dense Labelling. Given an image, produce per-pixel labels for that image. Labels include the distance of that pixel to the camera (Nathan Silberman & Fergus, 2012), surface orientation (Bae et al., 2021) or semantic class (Lin et al., 2014).

Image Classification. Given an image and optionally a target bounding box, generate a class name or tag of that image or target region. This group includes image classification (Deng et al., 2009) and object categorization (Pinz et al., 2006) datasets.

Image Captioning. Given an image and optionally a bounding box, generate a natural language description of that image or target region. We include both crowd-sourced (Chen et al., 2015) and webly supervised (Changpinyo et al., 2021) captions.

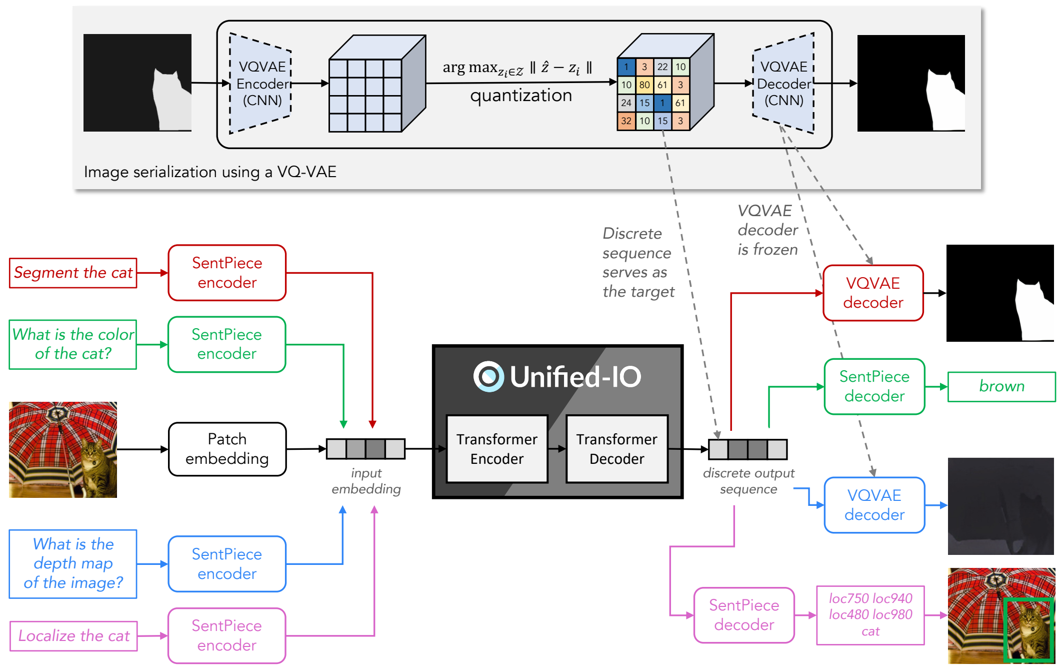

Figure 2: **Unified-IO.** A schematic of the model with four demonstrative tasks: object segmentation, visual question answering, depth estimation and object localization.

Vision & Language. A broad category for other tasks that require jointly reason over image content and a natural language query. There are many popular vision and language datasets, and we categories these datasets into 3 tasks – visual question answering (Antol et al., 2015); relationship detection (Lu et al., 2016) and grounded VQA (Chen et al., 2022a).

NLP. Tasks with text as the only input and output modalities, including text classification (Williams et al., 2018), question answering (Rajpurkar et al., 2016) and text summarization (Graff et al., 2003).

Language Modeling. The masking language modeling pre-training task (See Section 3.3) using text from C4 (Raffel et al., 2020) and Wikipedia (Foundation), which we include to ensure the knowledge gained from language pre-training is not lost during multi-task training. Other pre-training tasks are not included because the relevant datasets are already used in other supervised tasks (*e.g.*, for captioning or classification).

Table 1 shows the details of tasks and groups. We list an example dataset source, number of datasets, number of examples, percent of the total number of examples, and sampling rate during training (Section 3.3) for each group and task. Subsequent columns show what modalities are required for the inputs and outputs. We defer additional task details, inference details, the complete list of datasets and visualizations to the Appendix A.1.

3 Unified-IO

Our goal is to build a single unified model that can support a diverse set of tasks across computer vision and language with little to no need for task-specific customizations and parameters. Such unified architectures can be applied to new tasks with little to no knowledge of the underlying machinery, enable general pre-training to benefit many diverse downstream applications, be jointly trained on a large number of tasks, and better allows knowledge to be shared between tasks.

3.1 Unified Task Representations

Supporting a variety of modalities such as images, language, boxes, binary masks, segmentation masks, *etc.* without task-specific heads requires representing these modalities in a shared and unified space. To do this, we discretize the text, images, and other structured outputs in our tasks and represent them with tokens drawn from a unified and finite vocabulary.

Text representation. Following Raffel et al. (2020), text inputs and outputs are tokenized using SentencePiece (Kudo & Richardson, 2018). Following past works such as McCann et al. (2018); Raffel et al. (2020); Gupta et al. (2022a); Wang et al. (2022b) we also specify each task with a natural language prompt (excluding some tasks like VQA, which are fully specified by their text inputs) in order to indicate what task should be performed. For example, "*What is the depth map of the image?*" for depth estimation or "*What region does "cat" describe?*" for object localization.

Images and dense structures representation. A variety of tasks in computer vision requires the model to produce high-dimensional outputs such as images (*e.g.*, image in-painting) or per-pixel labels (*e.g.*, depth estimation). To handle these modalities, we first convert per-pixel labels into RGB images. For depth, we construct a grayscale image by normalizing the depth map. For surface normal estimation, we convert the $x/y/z$ orientations into $r/g/b$ values. For segmentation, we map each instance present in the image to a unique color. We randomly select colors for each instance and specify the color-to-class mapping in the text instead of using universal color-to-class mapping. This avoids requiring a fixed list of classes and avoids having colors that may only be marginally different due to the presence of a large number of classes.

Then we encode these images as discrete tokens using a VQ-GAN. In particular, we use the imagenet-pretrained VQ-GAN from Esser et al. (2021) with 256×256 resolution, compression ratio of 16, and 16384 codebook size. The VQ-GAN codebook is added to the vocabulary as additional tokens that can be generated by the decoder. During training, the tokens for the target image are used as targets. During inference, the VQ-GAN decoder is used to convert the generated image tokens into an output image.

Sparse structures representation. We encode sparse structures such as bounding boxes or human joints by adding 1000 special tokens to the vocabulary to represent discretized image coordinates (Chen et al., 2022b). Points are then encoded with a sequence of two such tokens, one for the x and one for the y coordinates, and boxes are encoded using a sequence of four tokens, two for the upper right corner and two for the lower left corner. Labeled boxes are encoded as a box followed by a text class label, and joints are encoded as a sequence of points followed by a text visibility label. This allows us to handle a wide variety of tasks that use these elements in their inputs or output (see Appendix A.1 for examples).

3.2 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 input images, we reshape the image into a sequence of patches that are embedded with linear projection similar to Dosovitskiy et al. (2021). Second, we expand the vocabulary to include the location tokens and the 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 essential to image tasks.

We use a maximum of 256 and 128 text tokens for inputs and outputs respectively, and a maximum length of 576 (*i.e.* 24×24 patch encoding from a 384×384 image) for image inputs and 256 (*i.e.* 16×16 latent codes from a 256×256 image) for image outputs. In this work, we present four versions of UNIFIED-IO ranging from 71 million to 2.9 billion parameters, as detailed in Table 2.

3.3 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

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

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., 2020; Wang et al., 2022b) 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 image. When another modality is present, *i.e.* image or text, the model can use information from that modality to complete the tasks.

We construct the pre-training dataset by incorporating publicly available language data (i.e., plain texts from Common Crawl), 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., 2022d).

We pre-train UNIFIED-IO on this combination of datasets with an in-batch mixing strategy. We equally sample data with the text and image denoising objective. For text denoising, half of the samples are from pure text data, *i.e.* C4 and Wikipedia. The other half is constructed from image and class data, such as Imagenet21k (Ridnik et al., 2021) or image and caption data, such as YFCC15M (Radford et al., 2021). For image denoising, we also use the same caption and class data and some image-only data from datasets for our vision tasks. We sample from datasets in proportion to dataset size. See Appendix A.2 for details.

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 95 datasets from 62 publicly available data sources. See Section 2 for task details and Appendix A.1 for dataset visualizations.

We jointly train UNIFIED-IO on this large set of datasets by mixing examples from these datasets within each batch. We equally sample each group ($1/8$) except for image synthesis ($3/16$) and dense labeling ($1/16$) since dense labeling has significantly fewer data and image synthesis has significantly more data than other groups. Within each group, we sample datasets proportional to the square root of their size to better expose the model to underrepresented tasks. Due to the large variance in dataset size, some tasks are still rarely sampled (*e.g.* depth estimation only has a 0.43% chance of being sampled). See Appendix A.3 for details and visualizations.

Implementation Details. Due to space limitation, see Appendix A.4 for implementation details.

4 EXPERIMENTS

We now present results for UNIFIED-IO on the GRIT benchmark (Sec 4.1), ablate training data via the GRIT ablation benchmark (Sec 4.2) and evaluate UNIFIED-IO on 16 other benchmarks in computer vision and NLP (Sec 4.3). Appendix A.5 shows evaluation on same concept and new concept on GRIT and A.6 shows the prompt generalization. Qualitative examples are in A.7.

4.1 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 has seven tasks that cover a range of visual skills with varying input and output modalities and formats: categorization, localization, VQA, refer expression, segmentation, keypoint, and surface normal estimation.

		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 [4]	-	-	-	-	-	-	-	-	-	-	-	-	**49.6**	**50.5**	7.2	7.1
1	Mask R-CNN [41]	-	-	44.7	45.1	-	-	-	-	26.2	26.2	**70.8**	**70.6**	-	-	20.2	20.3
2	GPV-1 [38]	33.2	33.2	42.8	42.7	50.6	49.8	25.8	26.8	-	-	-	-	-	-	21.8	21.8
3	CLIP [86]	48.1	-	-	-	-	-	-	-	-	-	-	-	-	-	6.9	-
4	OFA$_{\text{LARGE}}$ [107]	22.6	-	-	-	72.4	-	61.7	-	-	-	-	-	-	-	22.4	-
5	GPV-2 [52]	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	-	60.7	-
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 3: 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. Results of CLIP, OFA obtained from GRIT challenge.

UNIFIED-IO is the first model to support all seven tasks in GRIT. As seen in Table 3, UNIFIED-IO$_{\text{XL}}$ outperforms all prior submissions to GRIT obtaining average accuracy of 64.3 on test. The next best submission is GPV-2 (Kamath et al., 2022) which obtains 32.0 and can only support 4 out of 7 tasks. UNIFIED-IO$_{\text{XL}}$ also outperforms the multi-task checkpoint of OFA$_{\text{LARGE}}$ (Wang et al., 2022b) on VQA, refer expression and categorization.

Mask R-CNN (He et al., 2017) is a strong baseline for core vision tasks. UNIFIED-IO$_{\text{XL}}$ outperforms Mask R-CNN on localization and segmentation. The reason is UNIFIED-IO$_{\text{XL}}$ shows little degradation in performance between same concept and new concept as discussed in Appendix A.5. On keypoint, our model is worse compared to Mask R-CNN (68.1 *vs*. 70.8). The reason is we have 2-stage inference for keypoint – first locate the person using the object localization prompt, then find keypoints for each detected region.

NLL-AngMF (Bae et al., 2021) is a SOTA model for surface normal estimation. Our model gets strong results compared to NLL-AngMF (44.3 *vs*. 49.6). Since our image tokenizer is only pre-trained on ImageNet without any surface normal data, the upper bound of our method through reconstruction is 59.8 on FrameNet (Kazemzadeh et al., 2014). This suggests our score could be considerably improved by training a stronger image tokenizer.

4.2 ABLATIONS

To better understand how multi-tasking affects learning, we perform ablations by leaving out individual task groups from multi-task training. Due to computational constraints, we ablate UNIFIED-IO$_{\text{LARGE}}$ and train for 250k steps. If ablating a task group, we reduce the number of training steps so that all models are trained on approximately the same number of examples for each of the remaining task groups. Results are shown in Table 4 on GRIT and MNLI (Williams et al., 2018).

In spite of supporting a large number of heterogeneous tasks, Unified-IO is able to perform well across all tasks. Reducing this heterogeneity by removing task groups does not impact the performance of individual tasks significantly. This is notable since removing a task group significantly reduces the scope of what a model needs to learn while keeping the model capacity fixed. This empirically demonstrates the effectiveness of the proposed unified architecture for massive heterogeneous task support.

An exception is that removing the NLP group significantly boosts categorization, which might indicate that the sentence classification task interferes with image classification. Removing captioning also boosts performances on VQA and a few other tasks, which might be caused by captioning requiring a relatively large amount of model capacity to learn free-form text generation, in contrast to VQA that requires short answer phrases from a limited vocabulary. Removing image synthesis causes a major regression in keypoint. Manual inspection shows that the model predicts standing-human shaped keypoints even for people in very different postures, suggesting the model learned to rely on priors instead of the image content. We also see minor regressions in localization and referring expression, suggesting that image synthesis tasks, possibly image in-painting in particular, had a surprising positive transfer to understanding sparse structured outputs. It is possible that an ablation analysis on the XL model may yield different outcomes, but we are unable to perform an XL-based analysis due to limited compute.

Model	Step	Categorization	Localization	VQA	Refexp	Segmentation	Keypoint	Normal	MNLI
UNIFIED-IO$_{\text{LARGE}}$	250k	50.3	63.4	65.7	73.4	51.8	69.2	40.7	85.1
w/o Image Synthesis	200k	52.7	62.9	64.2	72.0	53.6	18.3	**42.2**	84.3
w/o Sparse	220k	52.6	-	64.1	-	51.3	-	38.5	83.8
w/o Dense	235k	49.5	62.4	65.6	72.9	-	66.7	-	84.8
w/o Classification	220k	-	63.1	64.0	73.7	52.1	66.8	39.1	84.6
w/o Captioning	220k	49.7	**65.0**	**68.0**	**74.7**	**54.2**	67.4	39.2	**85.3**
w/o V&L	220k	50.9	-	-	72.5	51.9	70.0	38.2	84.4
w/o NLP	220k	**56.1**	64.3	65.9	74.6	52.0	69.3	39.9	-
w/o Language Modelling	220k	52.9	64.7	66.7	**74.7**	52.7	**70.2**	39.9	83.5

Table 4: Ablation study on holding out tasks groups and evaluating on GRIT and MNLI (Williams et al., 2018)

	NYUv2	*ImageNet*	*Place365*	*VQAv2*	*OkVQA*	*A-OkVQA*	*VizWizQA*	*VizWizG*	*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-	BinsFormer	CoCa	MAE	CoCa	KAT	GPV2	Flamingo	MAC-Caps	JSL	OFA	SVT	CoCa	-	OFA	Turing NLR	ST-MOE	DeBERTa
tuned SOTA	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 5: 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., 2017), A-OKVQA (Schwenk et al., 2022), VizWizVQA (Gurari et al., 2018), VizWizG (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).

4.3 RESULTS ON ADDITIONAL TASKS

We report results on 16 additional tasks used in our 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 avoid extensive task-specific tricks like color jittering, horizontal flipping, CIDEr optimization, and label smoothing, which are often responsible for considerable gains in individual task performance. We leave such task-specific tuning for future work. See Table 5 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 SOTA (Li et al., 2022e) but ahead of the 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 the knowledge of many fine-grained classes despite being massively multi-tasked. Notably, we achieve this without the extensive data augmentations methods typically used by SOTA models (Yu et al., 2022a; 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 SOTA results on A-OKVQA. Relative to Flamingo, UNIFIED-IO performs better on VizWiz-QA but worse on OK-VQA.

Image Captioning. Despite the lack of CIDEr optimization, 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 SOTA models (Smith et al., 2022; Zoph et al., 2022; He et al., 2021). This can partly be attributed to scale. Modern NLP models contain 100 billion+ parameters and with more extensive NLP pre-training.

4.4 LIMITATIONS

For object detection, while UNIFIED-IO generally produces accurate outputs (see Appendix A.7), we find the recall is often poor in cluttered images. Prior work (Chen et al., 2022b) has shown this can be overcome with extensive data augmentation techniques, but these methods are not currently integrated into UNIFIED-IO. Our use of a pre-trained VQ-GAN greatly simplifies our training and is surprisingly effective for dense prediction tasks. However, it does mean UNIFIED-IO has limited image generation capabilities (recent works (Yu et al., 2022b) have shown this method can be greatly improved but was not available at the time of development). We also found in a small-scale study that our model does not always understand prompts not in the training data (see Appendix A.6).

5 RELATED WORK

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 (He et al., 2017; Liu et al., 2019; Lu et al., 2020). However, this requires manually designing a specialized head for each task and potentially limits transfer across tasks. An alternative is to build *unified* models – models that can complete many different tasks without task-specialized components. In NLP, 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).

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. Many models have been proposed for tasks with text and/or image input and text output (Cho et al., 2021; Wang et al., 2022d; Li et al., 2022b; Wang et al., 2021; Kaiser et al., 2017; Sun et al., 2022; Chen et al., 2022d; Wang et al., 2022c). However, these models can not produce any structured or visual output.

More recent unified models can additionally support image locations, which allows tasks like object detection or region captioning. This can be done by using bounding boxes proposed by an object detector (Cho et al., 2021; Kamath et al., 2022) or including a bounding box output head (Gupta et al., 2022a; Dou et al., 2022; Chen et al., 2022c; Kamath et al., 2021; Li et al., 2022d). Alternatively, image locations can be encoded as special tokens in the input/output text (Yang et al., 2021; Yao et al., 2022; Zhu et al., 2022) following Chen et al. (2022b). UNIFIED-IO follows this design, but applies it to a wider set of tasks than previous works, including key-point estimation, image in-painting, and region captioning.

Concurrent to our work, OFA (Wang et al., 2022b) proposes a similar approach that also supports image locations and text-to-image synthesis. However, OFA does not support dense labeling tasks such as depth estimation, segmentation, and surface normal estimation. Other closely related models include UViM (Kolesnikov et al., 2022) which generates a discrete guiding code for a D-VAE to build an autoregressive model for panoptic segmentation, depth prediction, and colorization, and Pix2Seq v2 (Chen et al., 2022c) which extends Pix2Seq to segmentation, keypoint estimation, and image captioning. UNIFIED-IO covers all these tasks and more, and focuses on multi-tasking rather then task-specific fine-tuning. Due to space limits, additional discussions are presented in Appendix A.8.

6 CONCLUSION

We have presented UNIFIED-IO, a unified architecture that supports a large variety of computer vision and NLP tasks with diverse inputs and outputs, including images, continuous 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. The 2.9B parameter UNIFIED-IO XL model is jointly trained on 90+ datasets, 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.

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 *ICCV*, 2019. 8

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. 8, 26

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015. 4

Gwangbin Bae, Ignas Budvytis, and Roberto Cipolla. Estimating and exploiting the aleatoric uncertainty in surface normal estimation. In *ICCV*, 2021. 3, 7

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

Roy Bar-Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. The second pascal recognising textual entailment challenge. In *Proceedings of the second PASCAL challenges workshop on recognising textual entailment*, volume 6, pp. 6–4. Venice, 2006. 22

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. The fifth pascal recognizing textual entailment challenge. In *TAC*, 2009. 22

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. The SNLI corpus. 2015. 22

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *NeurIPS*, 2020. 1, 9

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*, 2017. 22

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. 3, 17

Chongyan Chen, Samreen Anjum, and Danna Gurari. Grounding answers for visual questions asked by visually impaired people. In *CVPR*, 2022a. 4, 8, 21, 24

Ting Chen, Saurabh Saxena, Lala Li, David J Fleet, and Geoffrey Hinton. Pix2seq: A language modeling framework for object detection. In *ICLR*, 2022b. 5, 9, 18

Ting Chen, Saurabh Saxena, Lala Li, Tsung-Yi Lin, David J Fleet, and Geoffrey Hinton. A unified sequence interface for vision tasks. *arXiv preprint arXiv:2206.07669*, 2022c. 9

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022d. 9

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. 3, 8, 17

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *ICML*, 2021. 2, 9

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 preprint arXiv:2204.02311*, 2022. 9

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *NAACL*, 2019. 8, 22

Marcella Cornia, Lorenzo Baraldi, Giuseppe Fiameni, and Rita Cucchiara. Universal captioner: long-tail vision-and-language model training through content-style separation. *arXiv*, 2021. 8

Ido Dagan, Oren Glickman, and Bernardo Magnini. The pascal recognising textual entailment challenge. In *Machine Learning Challenges Workshop*, pp. 177–190. Springer, 2005. 22

Marie-Catherine De Marneff, Mandy Simons, and Judith Tonhauser. The commitmentbank: Investigating projection in naturally occurring discourse. *proceedings of Sinn und Bedeutung 23*, 2019. 22

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. 3, 8, 20

Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. RedCaps: Web-curated image-text data created by the people, for the people. *arXiv*, 2021. 3, 17

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019. 5, 22, 26

William B. Dolan and Chris Brockett. Automatically constructing a corpus of sentential paraphrases. In *IWP*, 2005. 8, 22

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

Zi-Yi Dou, Aishwarya Kamath, Zhe Gan, Pengchuan Zhang, Jianfeng Wang, Linjie Li, Zicheng Liu, Ce Liu, Yann LeCun, Nanyun Peng, et al. Coarse-to-fine vision-language pre-training with fusion in the backbone. *arXiv preprint arXiv:2206.07643*, 2022. 9

Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. Searchqa: A new q&a dataset augmented with context from a search engine. *arXiv preprint arXiv:1704.05179*, 2017. 21

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021. 2, 5, 6

Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. Mrqa 2019 shared task: Evaluating generalization in reading comprehension. *arXiv preprint arXiv:1910.09753*, 2019. 21

Wikimedia Foundation. Wikimedia downloads. URL `https://dumps.wikimedia.org`. 4, 22

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. The third pascal recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL workshop on textual entailment and paraphrasing*, pp. 1–9. Association for Computational Linguistics, 2007. 22

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. In *CVPR*, 2017. 8, 21

David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. English gigaword. *Linguistic Data Consortium, Philadelphia*, 4(1):34, 2003. 4, 22

Liangke Gui, Borui Wang, Qiuyuan Huang, Alex Hauptmann, Yonatan Bisk, and Jianfeng Gao. Kat: A knowledge augmented transformer for vision-and-language. *arXiv*, 2021. 8

Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019. 3, 17

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, 5, 7, 9

Tanmay Gupta, Ryan Marten, Aniruddha Kembhavi, and Derek Hoiem. GRIT: General robust image task benchmark. *arXiv*, 2022b. 2, 6

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 *CVPR*, 2018. 8, 21

Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017. 1, 7, 9

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 6, 8

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. DeBERTa: Decoding-enhanced bert with disentangled attention. In *ICLR*, 2021. 9

Ronghang Hu and Amanpreet Singh. Unit: Multimodal multitask learning with a unified transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1439–1449, 2021. 2

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. 19

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Cosmos qa: Machine reading comprehension with contextual commonsense reasoning. *arXiv*, 2019b. 22

Shankar Iyer, Nikhil Dandekar, and Kornel Csernai. First quora dataset release: Question pairs, 2017. URL https://data.quora.com/First-Quora-Dataset-Release-Question-Pairs. 22

Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier J Henaff, Matthew Botvinick, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. Perceiver io: A general architecture for structured inputs & outputs. In *ICLR*, 2022. 26

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147. URL https://aclanthology.org/P17-1147. 21

Lukasz Kaiser, Aidan N Gomez, Noam Shazeer, Ashish Vaswani, Niki Parmar, Llion Jones, and Jakob Uszkoreit. One model to learn them all. *arXiv preprint arXiv:1706.05137*, 2017. 9

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1780–1790, 2021. 9

Amita Kamath, Christopher Clark, Tanmay Gupta, Eric Kolve, Derek Hoiem, and Aniruddha Kembhavi. Webly supervised concept expansion for general purpose vision models. *arXiv*, 2022. 7, 8, 9

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, 2014. 3, 7, 18

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. Looking beyond the surface:a challenge set for reading comprehension over multiple sentences. In *Proceedings of North American Chapter of the Association for Computational Linguistics (NAACL)*, 2018. 22

Tushar Khot, Ashish Sabharwal, and Peter Clark. Scitail: A textual entailment dataset from science question answering. In *AAAI*, 2018. 8, 22

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*, 2022. 8, 9, 24

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2017. 3, 17, 20, 21

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. 3, 17, 21

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026. 21

Hector Levesque, Ernest Davis, and Leora Morgenstern. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*, 2012. 22

Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, Hehong Chen, Guohai Xu, Zheng Cao, et al. mplug: Effective and efficient vision-language learning by cross-modal skip-connections. *arXiv preprint arXiv:2205.12005*, 2022a. 26

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 26

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022b. 9, 26

Linjie Li, Zhe Gan, Kevin Lin, Chung-Ching Lin, Zicheng Liu, Ce Liu, and Lijuan Wang. Lavender: Unifying video-language understanding as masked language modeling. *arXiv preprint arXiv:2206.07160*, 2022c. 26

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, 26

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10965–10975, 2022d. 9

Zhenyu Li, Xuyang Wang, Xianming Liu, and Junjun Jiang. Binsformer: Revisiting adaptive bins for monocular depth estimation. *arXiv*, 2022e. 8

Paul Pu Liang, Yiwei Lyu, Xiang Fan, Shengtong Mo, Dani Yogatama, Louis-Philippe Morency, and Ruslan Salakhutdinov. Highmmt: Towards modality and task generalization for high-modality representation learning. *arXiv preprint arXiv:2203.01311*, 2022. 26

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. 3, 17, 18

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4487–4496. Association for Computational Linguistics, 2019. URL `https://www.aclweb.org/anthology/P19-1441`. 9

Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *European Conference on Computer Vision*, 2016. 4

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 2019. 2, 26

Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, D. Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *CVPR*, 2020. 6, 9

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P11-1015`. 22

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, 2016. 18

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 21

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. 22

Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from RGBD images. In *ECCV*, 2012. 3, 8, 19, 24

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. 8, 21

Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *ArXiv*, abs/2208.06366, 2022. 26

Mohammad Taher Pilehvar and os'e Camacho-Collados. Wic: 10, 000 example pairs for evaluating context-sensitive representations. *CoRR*, abs/1808.09121, 2018. URL http://arxiv.org/abs/1808.09121. 22

Axel Pinz et al. Object categorization. *Foundations and Trends® in Computer Graphics and Vision*, 1(4): 255–353, 2006. 3

Sarah Pratt, Mark Yatskar, Luca Weihs, Ali Farhadi, and Aniruddha Kembhavi. Grounded situation recognition. In *ECCV*, 2020. 8

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 6, 7, 23, 26

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. 1, 2, 4, 5, 6, 9, 21, 22, 24

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. In *EMNLP*, 2016. 4, 21

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv*, 2022. 17

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. 26

Anthony D Rhodes, Max H Quinn, and Melanie Mitchell. Fast on-line kernel density estimation for active object localization. In *2017 international joint conference on neural networks (IJCNN)*, pp. 454–462. IEEE, 2017. 3

Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv*, 2021. 6, 20, 23

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*, 2011. 22

Alexander M. Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractive sentence summarization. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015. doi: 10.18653/v1/d15-1044. URL http://dx.doi.org/10.18653/v1/D15-1044. 22

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. *arXiv*, 2022. 8, 21

Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pp. 4596–4604. PMLR, 2018. 25

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15638–15650, 2022. 26

Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, Elton Zheng, Rewon Child, Reza Yazdani Aminabadi, Julie Bernauer, Xia Song, Mohammad Shoeybi, Yuxiong He, Michael Houston, Saurabh Tiwary, and Bryan Catanzaro. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. *arXiv*, 2022. 9

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1631–1642, 2013. 22

Benyuan Sun, Jin Dai, Zihao Liang, Congying Liu, Yi Yang, and Bo Bai. Gppf: A general perception pre-training framework via sparsely activated multi-task learning. *arXiv preprint arXiv:2208.02148*, 2022. 9, 26

Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *EMNLP*, 2019. 2, 26

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pp. 191–200, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-2623. URL `https://aclanthology.org/W17-2623`. 21

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. 20

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. 22

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. 21, 22

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022a. 26

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*, 2022b. 2, 5, 6, 7, 9, 21, 26

Wenhui Wang, Hangbo Bao, Li Dong, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *arXiv preprint arXiv:2111.02358*, 2021. 9

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022c. 9, 26

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*, 2022d. 6, 9, 21, 26

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. Neural network acceptability judgments. *arXiv preprint arXiv:1805.12471*, 2018. 22

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. 20

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1112–1122. Association for Computational Linguistics, 2018. URL `http://aclweb.org/anthology/N18-1101`. 4, 7, 8, 22

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. 20

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment: A novel task for fine-grained image understanding. *arXiv*, 2019. 8, 21

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Faisal Ahmed, Zicheng Liu, Yumao Lu, and Lijuan Wang. Unitab: Unifying text and box outputs for grounded vision-language modeling. 2021. 9

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1259. URL `https://aclanthology.org/D18-1259`. 21

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. 19

Yuan Yao, Qianyu Chen, Ao Zhang, Wei Ji, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Pevl: Position-enhanced pre-training and prompt tuning for vision-language models. *arXiv preprint arXiv:2205.11169*, 2022. 9

Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. Situation recognition: Visual semantic role labeling for image understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5534–5542, 2016. 21

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*, 2022a. 8, 26

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022b. 9

Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 26

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6720–6731, 2019a. 21

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019b. 22

Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. *Advances in Neural Information Processing Systems*, 34:23634–23651, 2021. 26

Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. Merlot reserve: Neural script knowledge through vision and language and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16375–16387, 2022. 26

Yuan Zhang, Jason Baldridge, and Luheng He. PAWS: Paraphrase Adversaries from Word Scrambling. In *Proc. of NAACL*, 2019. 22

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *TPAMI*, 2017. 8, 20

Chaoyang Zhu, Yiyi Zhou, Yunhang Shen, Gen Luo, Xingjia Pan, Mingbao Lin, Chao Chen, Liujuan Cao, Xiaoshuai Sun, and Rongrong Ji. Seqtr: A simple yet universal network for visual grounding. *arXiv preprint arXiv:2203.16265*, 2022. 9

Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam Shazeer, and William Fedus. Designing effective sparse expert models. *arXiv*, 2022. 9