i-Code V2: An Autoregressive Generation Framework over Vision, Language, and Speech Data

Ziyi Yang, Mahmoud Khademi, Yichong Xu, Reid Pryzant, Yuwei Fang, Chenguang Zhu, Dongdong Chen, Yao Qian, Mei Gao, Yi-Ling Chen, Robert Gmyr, Naoyuki Kanda, Noel Codella, Bin Xiao, Yu Shi, Lu Yuan, Takuya Yoshioka, Michael Zeng, Xuedong Huang
Microsoft AI

Abstract

The convergence of text, visual, and audio data is crucial towards human-like artificial intelligence, however the current Vision-Language-Speech landscape is dominated by encoder-only models that lack generative abilities. We propose closing this gap with i-Code V2, the first model capable of generating natural language from any combination of Vision, Language, and Speech data. i-Code V2 leverages state-of-the-art single-modality encoders, combining their outputs with a new modality-fusing encoder to project combinations of modalities into a shared representational space. Language tokens are generated from these representations via an autoregressive decoder. i-Code V2 is pretrained end-to-end on a large collection of dual- and single-modality datasets with a novel text completion objective that can be generalized across arbitrary combinations of modalities. i-Code V2 matches or outperforms state-of-the-art single- and dual-modality baselines on 7 multimodal tasks, demonstrating the power of generative multimodal pretraining across a diversity of tasks and signals.

1 Introduction

Pretrained Large language models (LLMs) have experienced massive success as general-purpose solutions for multiple tasks (Brown et al., 2020). However, a large gap persists between the capabilities of LLMs and true humanlike intelligence. This is partially because humans perceive a variety of sensory inputs while LLMs are typically restricted to Language (L) data and unable to understand or generalize to other modalities such as Vision (V) and Speech audio (S).

Recently, the field of multimodal AI, which aims to develop AI systems capable of modeling multiple kinds of signals, has witnessed significant progress including new learning techniques (Radford et al., 2021; Bao et al., 2022; Alayrac et al., 2022), training data (Schuhmann et al., 2022; Zellers et al., 2022a; Yang et al., 2023), and model architectures (Su et al., 2020; Li et al., 2019; Xu et al., 2022).

Despite this progress in multimodal AI, most research has focused on understanding pairs of modalities, such as speech-language and vision-language, and the fast-growing subfield of triple-modality AI (Language, Vision, Speech) remains limited to encoder-only models (Akbari et al., 2021; Zellers et al., 2022b; Yang et al., 2023). This paper proposes i-Code V2, one of the the first encoder-decoder generative models for the triple-modality setting. i-Code V2 can flexibly generate text from arbitrary combinations of Language, Vision, and Speech data. This model addresses three ongoing challenges within multimodal research.

First, most existing vision-language-speech models are encoder-only, i.e. they can conduct discriminative tasks such as multimodal classification but not generative ones like visual question answering or automatic speech recognition. i-Code V2 enables the model to generate content from multimodal signals, unlocking more diverse applications and improved discriminative performance.

Second, most existing triple-modality research leverages triple-modality data (i.e. video with subtitles and audio track). However, the three modalities in video data can be noisily aligned (Miech et al., 2019) which degrades downstream pretraining. Furthermore, the available high quality video data is several orders of magnitudes smaller in size than single- or dual-modality ones. E.g., the largest publicly available image-caption dataset LAION (Schuhmann et al., 2022) has 5 billion pairs (335 billion text tokens) while the largest video dataset MERLOT has 180M videos (5 billion text tokens) (Akbari et al., 2021; Zellers et al., 2022a). i-Code V2 proposes a novel method for efficiently leveraging these larger and higher-quality dual- and single-modality datasets within a triple-modality pretraining framework. We accomplish this with

Findings of the Association for Computational Linguistics: NAACL 2024, pages 1615–1627

a new, generalized sequence-to-sequence pretraining objective which unifies assorted multimodal objectives into simple text completion.

Third, multimodal tasks are diverse in settings and data formats, e.g. Automatic Speech Recognition (ASR), vision QA, sentiment analysis, etc. Existing techniques apply separate inference strategies to each problem type, adding complexity and overhead for practitioners. i-Code V2 unifies all tasks under its text completion framework, rendering multimodal inference and cross-task transfer easier for practitioners.

i-Code V2 is built on top of state-of-the-art single-modality models: the vision and speech modalities are encoded with single-modality encoder respectively. Then encoded features and text token embeddings are inputted to a joint vision-language-speech encoder, which merges the different modalities into a shared representational space. Last, a language decoder, conditioned on the joint encoder via a cross-attention mechanism, is trained to generate language tokens autoregressively.

We evaluate i-Code V2 on 7 datasets: multimodal summarization, multimodal dialogue generation, multimodal sentiment analysis, vision QA, vision captioning, and ASR. Notably, i-Code V2 outperforms previous SOTA models on MSMO (multimodal summarization), Image Chat (multimodal dialogue generation), UR-FUNNY (multimodal sentiment analysis). i-Code V2 also exhibits competitive performance compared to specialized dual-modality models on vision QA, vision captioning, and ASR, suggesting the power of integrative multimodal pretraining.

In summary, our key contributions are threefold:

1. We propose i-Code V2, one of the first vision-language-speech generative models that can generate natural language from one-, two- or three-modality inputs of image, video, language and speech.

2. We propose a novel multimodal generative pretraining framework using large-scale uni- and dual-modality datasets with a novel cross-modality text completion framework. Utilizing a sequence-to-sequence objective, instead of modality-specific objectives, enables flexible application to various training goals and streamlines in training and inference.

3. i-Code V2 shows SOTA or competitive performance across several multimodal tasks and domains, including multimodal summarization and dialogue generation, and video sentiment analysis.

2 Related Work

Multimodal Learning studies extracting and incorporating information from vision, language, and speech modalities. A recent advance is unifying models of different modalities to the transformer. For example, representing vision and language with one multimodal transformer model has shown great performance in image caption (Wang et al., 2022a; Alayrac et al., 2022), vision classification (Yu et al., 2022), vision question answering (Yu et al., 2022; Li et al., 2022), etc. Extracted image features (Chen et al., 2020) or projections of image patches (Wang et al., 2022e; Yu et al., 2022) are fused together with text token embeddings, then input to a multimodal encoder to obtain unified representations for vision and language. For vision-language-speech models, the multimodal encoder is pretrained on video data (Zellers et al., 2022a; Yang et al., 2023) or dual-modality data pairs (Yang et al., 2023). Multimodal representations can be integrated by a late-stage multimodal fusion network (Yang et al., 2023), or integrated early at the input stage (Zellers et al., 2022a).

Generative Multimodal Model can generate one modality from another modality or a combination of input modalities. E.g., image captioning (Johnson et al., 2016; Wang et al., 2022a), automatic speech recognition (Yu and Deng, 2016; Radford et al., 2022), text-to-image generation (Ramesh et al., 2021, 2022; Saharia et al., 2022; Rombach et al., 2022), etc. Several recent works propose to unify vision-language tasks with one homogeneous model architectures and schemes. For example, Wang et al. (2022c); Lu et al. (2022); Tang et al. (2022) unite vision and vision-language tasks, such as image classification, object detection, semantic segmentation, visual QA, document understanding, image generation, etc. Huang et al. (2023) recently proposed the Kosmos-1 model to generate text based on vision and text input.

Distinct from previous works, i-Code V2 can not only encode and merge vision, language, and speech modalities, but also generate natural language. It unifies various tasks across multimodal summarization, multimodal sentiment analysis, speech recognition, visual QA, and caption.

3 An Integrative Multimodal Generative Model

3.1 Model Architecture

i-Code V2 model consists of multimodal encoders and a language decoder. Following the spirit of integrative AI (Yang et al., 2023), the language, vision and speech modalities are encoded by their corresponding encoder or converted to numerical representations respectively, before being fused with each other. Leveraging pretrained models enables us to utilize the state-of-the-art model architecture for each modality. It is also computationally efficient since these models have already been extensively trained on single-modality data. This also gives us the flexibility of choosing preferred encoders. For example, we can use a medical-domain specific language encoder-decoder; or choose a smaller speech/vision encoder on devices having limited computation resources, without having to re-design the framework. We leverage the following state-of-the-art single-modality encoders:

Vision Encoder. In different multimodal scenarios, the vision modality can be either a single image or a video (especially when the input contains speech). To flexibly encode and represent the vision modality input, we opt to use OmniVL, a foundation model for both image-language and video-language (Wang et al., 2022b). It uses independent 2D/3D convolution-based patch tokenizers to first process image/video and a unified vision transformer to generate vision representations. It has 122 million parameters.

Speech Encoder. We use WavLM large (Chen et al., 2022), a speech encoder pretrained on 94k-hour data in a self-supervised manner. Pretraining objectives include masked speech denoising and predicting. The model architecture is a transformer encoder with Gated Relative Position Bias on top of a temporal CNN-based featurizer. The parameter size is 315 million.

Joint Vision-Language-Speech Encoder. We use a 24-layer transformer encoder to jointly encode vision, language, and speech modalities. After the vision and speech modality inputs are encoded by their respective encoder, a 1-layer projection (one for each modality) transforms the features into the same dimension as the text vocabulary embedding. Transformed features are concatenated with the text tokens embeddings and then input into the transformer layers for both inter- and intra-modality attention.

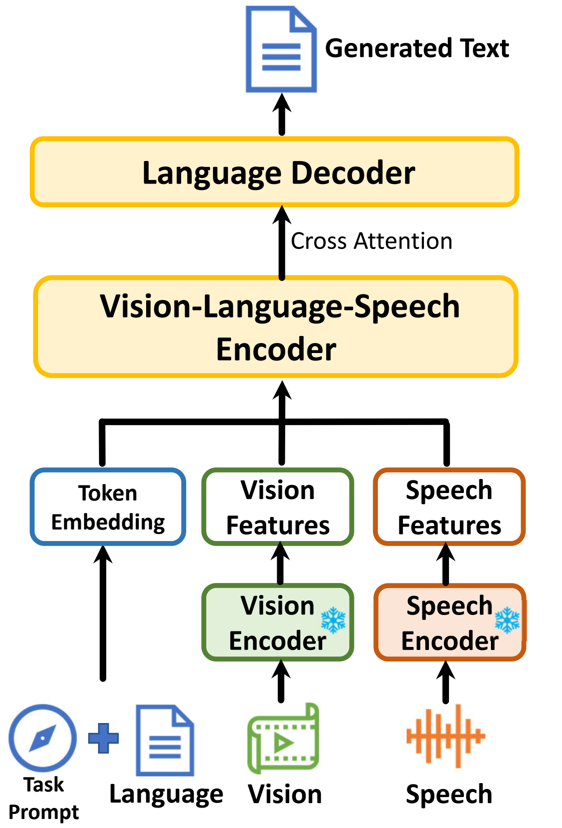

Figure 1: i-Code V2 Model Architecture. Parameters of vision and speech encoders are frozen during pretraining and are updated in finetuning.

We initialize the transformer layers of the joint encoder using the encoder part of the recently developed Z-Code++ summarization model, which has 485 million parameters and was pretrained using generative training objectives on 160G of English text data (He et al., 2022).

Language Decoder with Multimodal Cross-Attention. i-Code V2 then uses a decoder to generate textual sequences from the multimodal encoder output. The 24-layer decoder cross-attends with the multimodal representation from the joint Vision-Language-Speech encoder. We use the pretrained transformer *decoder* from the Z-Code++ model (485 million parameters) to initialize these parameters.

3.2 Large-Scale Multimodal Generative Pretraining

We leverage a collection of large-scale dual modality datasets to conduct speech-language generative pretraining, vision-language generative pretraining, and language-language generative pretraining. In particular, our pretraining objectives adopt a simple sequence-to-sequence strategy, which poses each modality-specific and cross-modality objective as a text completion. The multimodal pretraining process, task, and textual instructions are illustrated in Figure 2.

3.2.1 Vision-Language Generative Pretraining

Image Captioning. Given an image, the model predicts the corresponding textual caption. We use the 72.8 million subset of Florence image-text pair dataset (Yuan et al., 2021). The task prompt is "Generate the caption for this image: ".

Video Captioning. The pretraining task is to generate the caption of a video clip. We use the largest-scale publicly available video captioning dataset WebVid-10M (Bain et al., 2021), which contains 10.7M video-caption pairs. The task prompt is "Generate the caption for this video: ".

Vision Question & Answering. For this task, we use the VQA v2 training set, an open-ended vision question answering dataset (Antol et al., 2015), which has 443,757 question-answer pairs. The task prompt is "Answer the following question based on the image: ".

Vision-Augmented Text Reconstruction. This pretraining task aims to improve the model's ability on cross-modal understanding. We mask spans of the textual image caption and replace them with sentinel tokens, like T5 pretraining (Raffel et al., 2020). The model needs to predict masked out text spans, given the masked textual input and the image. The data resource is the same as in "Image Captioning". The task prompt is "Reconstruct the following text based on the image: ".

3.2.2 Speech-Language Generative Pretraining

We leverage the following labeled data for generative speech-language pretraining:

Speech transcription. This dataset contains 75k-hour human-transcribed speech utterances (Yang et al., 2023), collected from scenarios such as call center and AI voice assistant. The input is the speech utterance, and the target output is the transcription. The pretraining loss is the cross entropy between the target and prediction. The task prompt is "Transcribe the speech utterance to text: ".

Speech Sentiment Analysis. The goal of this task is to predict the sentiment of a speech utterance, e.g., from "highly negative" to "highly positive". We gather data from CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) (Zadeh et al., 2018) and Spoken Language Understanding Evaluation (SLUE) (Shon et al., 2022). The task prompt is "Predict the sentiment of this segment: ". The output target is the textual sequence of the "sentiment".

Speech Emotion Recognition. The task is to predict the emotion category of a speech utterance, including {happiness, sadness, anger, fear, disgust, surprise}. The dataset is from the emotion intensity subtask of CMU-MOSEI. The target generation sequence is the emotion category name. The task prompt is "Predict the emotion of this segment: ".

Speech-Augmented Text Reconstruction. Similar to "Vision-Augmented Text Reconstruction", we mask spans of the speech transcription and ask the model to predict masked-out text spans, given the speech input as well. The task prompt is "Reconstruct the following text based on the speech: ".

3.2.3 Language-only Generative Pretraining.

We include two high-quality text-only corpora, i.e., English Wikipedia and BookCorpus (Zhu et al., 2015), in pretraining as a supplement to the language-modality data of vision-language and speech-language datasets. This language-only pretraining task follows T5 where the input is span-masked text, and the output is the original masked span. The task prompt is "Reconstruct masked spans in the following text: ".

3.2.4 Pretraining Details

To expedite the pretraining process, we freeze the weights of speech and vision encoders, only updating the parameters of the Vision-Language-Speech encoder and the language decoder. For each optimization step, we select the pretraining dataset from the candidate pool using "Exponentially Smoothed Weighting (ESW)" sampling. ESW is widely used in multilingual pretraining (Devlin et al., 2019) where multilingual corpus sizes can be different with several magnitudes. Assume the size ratio of the dataset A in the overall training datasets is $P(A)$. We exponentiate the ratio by the factor $S < 1$ then we sample datasets according

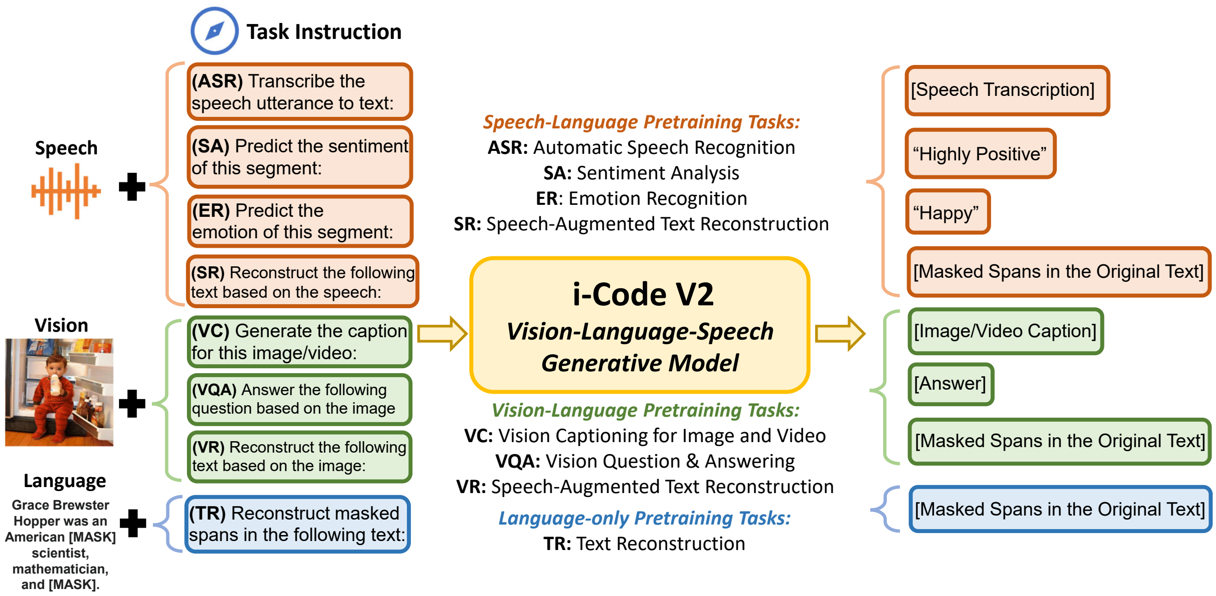

Figure 2: i-Code V2 multimodal pretraining. It unifies tasks across vision, language and speech domains to text completion/generation training objectives. Pretraining tasks include both unimodal, e.g. TR, and dual-modal ones, e.g., ASR, SA, ER, SR, VC, VQA, and VR (full names of task initials are provided in the figure).

the re-normalized exponential ratio $\frac{p(A)^S}{\sum_A p(A)^s}$. We use $S = 0.5$ in our setting.

We pretrain the i-Code V2 model on datasets introduced above for 1 epoch on 24 A100 GPUs, with batch size 8 (per-GPU) and three gradient accumulation steps. Having accumulation steps > 1 also makes the effective optimization batch contain data from different resources. We use AdamW (Loshchilov and Hutter, 2019) optimizer with starting learning rate 10^{-5}. The number of warm up steps is 2000, and the learning rate linearly decays to 5×10^{-6}.

4 Experiments

We test i-Code v2 on 7 datasets from assorted categories. In downstream tasks, we update parameters of Vision-Language-Speech encoder, language decoder, and single-modality encoders. Overall, i-Code V2 sets a new state-of-the-art in 3 tasks (MSMO, Image Chat, and UR-FUNNY) and remains highly competitive in the rest, suggesting the promise of integrative and generative multimodal pretraining.

4.1 Multimodal Summarization

We first evaluate the multimodal summarization task. Well studied in the field of natural language processing, in traditional summarization, input only contains language. However, in many real-world scenarios, such as multimedia coverage and online news article, key information is also included other modalities, e.g., pictures. We test i-Code V2 on the multimodal news summarization dataset MSMO (Zhu et al., 2018). We choose MSMO since the dataset is fully open-sourced, including the images in the article. Given a news article with image(s), the task is to generate a few-sentence summarization. Its training/validation/test split contains 293,965/10,355/10,261 news articles with images from Daily Mail website. The ground-truth "golden" summary is the highlight written by the news editor. The evaluation metrics are ROUGE scores (Lin, 2004). Baseline models include: **text-only summarization model**, e.g., BertSum (Liu and Lapata, 2019) model variants BertAbs and BertExtAbs, BART (Lewis et al., 2020), ZCode++ (He et al., 2022); **UniMS (Zhang et al., 2022)**, an encoder-decoder multimodal summarization model that can process multimodal inputs and select images; **MOF (Zhu et al., 2020)**, a multimodal generation model with the guidance of multimodal reference; **ATG/ATL/HAN**, these are baselines from the original MSMO dataset paper, that Point Generator Network (See et al., 2017) attending with global vision features(ATG), attending with local vision features (ATL), and hierarchical attention with local features.

The task prompt used in i-Code V2 is "Summarize this article with the images: ". Our model has the flexibility of encoding several images using the

Model	R1	R2	RL
BertAbs	39.02	18.17	33.20
BertExtAbs	39.88	18.77	38.36
BART	41.83	19.83	39.74
ZCode++	42.19	20.03	37.2
UniMS	42.94	20.50	**40.96**
MOF (enc)	41.05	18.29	37.74
MOF (dec)	41.20	18.33	37.80
HAN	40.82	18.30	37.70
ATL	40.86	18.27	37.75
ATG	40.63	18.12	37.53
i-Code V2	**44.7±0.2**	**21.0±0.3**	37.7±0.2

Table 1: Results on the multimodal news summarization MSMO test set.

video encoder. As shown in Table 1, compared with baseline models, i-Code V2 has shown competitive performance on ROUGE-1 and ROUGE-L. Compared with the language encoder-decoder ZCode++, that i-Code V2 encoder-decoder is initialized from, i-Code V2 shows considerable improvement, which demonstrates the effectiveness of the proposed multimodal pretraining.

4.2 Multimodal Dialogue Generation

i-Code V2 also has the ability perceive contextual multimodal signals to generate textual response. We test on the multimodal open-domain dialogue dataset Image-Chat (Shuster et al., 2020), each data example includes an image; the dialogue history between two speakers A and B; and speaker style traits. The goal is to generate the next-round dialogue. Baselines include: **BlenderBot (Roller et al., 2020)**, a ChatBot model of 2.7 Billion parameters pretrained on 1.5B Reddit comment conversations; **Multi-Modal BlenderBot(Shuster et al., 2021)**, the multimodal version of BlenderBot that fuses vision features from ResNet/Faster-RCNN in the multimodal text generation; **2AMMC (Ju et al., 2019)**, a multimodal generative model that combines ResNet and text transformer; **DialoGPT (Zhang et al., 2019)**, a GPT model trained on 147 million social media dialogues.

We can conveniently guide the model to generate dialogue in the speaker style with prompt "Generate the response for the dialogue in {style type} style: ". For fair comparison, we do not include baseline model that co-trains on multiple multimodal dialogue datasets e.g., Shuster et al. (2020). The evaluation metric includes F1 and ROUGE-L. Table 2 shows that i-Code V2 has significantly outperformed previous baselines on both metrics.

Model	F1	RL
DialoGPT	6.2	5.2
2AMMC	9.3	11.0
BlenderBot	9.2	12.3
Multi-Modal BlenderBot	13.1	18.0
i-Code V2	**15.5±0.2**	**18.6±0.3**

Table 2: Results on the multimodal dialogue generation dataset Image Chat.

Model	Accuracy
ZCode++	75.4
MulT (Tsai et al., 2019)	70.55
MISA (Hazarika et al., 2020)	70.61
MultiBench (Liang et al., 2021)	66.7
BBFN (Han et al., 2021)	71.68
LMF (Liu et al., 2018)	67.53
TFN (Zadeh et al., 2017)	68.57
i-Code V2	**79.59±0.18**

Table 3: Prediction accuracy on UR-FUNNY dataset.

4.3 Video Multimodal Sentiment Analysis

We further evaluate i-Code V2 on multimodal sentiment analysis datasets. E.g., in UR-FUNNY (Hasan et al., 2019), a humor detection dataset, the input is a video, the audio of the video, and the text transcript. The task is to predict whether the immediate laughter will follow the clip. The dataset contains 5306/1313/1638 humor instances for train/validation/test split, and 5292/1313/1652 for the not humor instances. Although previous models approached this problem as binary classification, we finetune i-Code V2 to directly predict the target sequence "funny"/"unfunny", with task prompt "Predict the sentiment of this clip: ". We compare i-Code V2 with baselines that use all three-modality inputs. i-Code V2 outperforms previous models by large margins (Table 3). This shows that the multimodal encoder in i-Code V2 can effectively fuse signals of vision, language and speech modalities, and the decoder can successfully attend with the multimodal encoder outputs.

4.4 Automatic Speech Recognition

Automatic Speech Recognition (ASR) transfers human-spoken language into text. We evaluate on the classical ASR dataset LibriSpeech (Panay-

Model	WER(%)↓
wav2vec 2.0	2.0
WavLM Large	2.1
Whisper Large	2.7
Whisper Medium	4.12
S2T Transformer Large	3.2
i-Code V2	3.86±0.17

Table 4: Word Error Rate (WER) on LibriSpeech dataset test-clean split.

otov et al., 2015). We finetune i-Code V2 on LibriSpeech 960h training data and test on the test-clean split. We compare i-Code V2 with the following models: **WavLM (Chen et al., 2022)**, a transformer-based speech encoder that is pre-trained on audio data with self-supervised learning; **wav2vec 2.0 (Baevski et al., 2020)**, a speech representation model with CNN-Transformer architecture, pretrained with a contrastive self-supervised task on quantized speech representations; **S2T Transformer (Wang et al., 2021)**, a transformer-based speech-to-text model provided in the Fairseq (Ott et al., 2019) sequence modeling toolkit; **Whisper (Radford et al., 2022)**, a recently developed speech recognition system that is pretrained on 680K hours of labeled speech-text transcript with multitask-supervision.

The task prompt is "transcribe the speech utterance to text: ". Results in Table 4 show that i-Code V2 is capable of decoding speech signals to language with performance close to models specifically designed for the ASR task. Note that WavLM Large result presented in Table 4 is using Connectionist temporal classification (CTC) decoding on top of the speech encoder, which is specifically designed for ASR task. While the language decoding in i-Code V2 is for general purpose. We notice that i-Code degrades on ASR performance compared with WavLM Large. It is worth pointing out that WavLM-large performance is obtained by adding a CTC component that is specifically designed for speech transcription. In contrast, i-Code V2 uses the general transformer decoder layer mechanism to generate text tokens. This can cause the performance discrepancy. Moreover, different modalities can be competing for the modeling capacity. Increasing the language decoder size is a potential solution.

Model	Accuracy
Closed-Vocabulary	
VisualBERT (Li et al., 2020)	71.0
LXMERT (Tan and Bansal, 2019)	72.5
FLAVA	72.8
OSCAR	73.16
VL-BERT (Su et al., 2020)	72.2
BLIP (Li et al., 2022)	78.32
CoCa (Yu et al., 2022)	**82.3**
Open-Vocabulary	
Flamingo*(Alayrac et al., 2022)	82.1
i-Code V2	75.10

Table 5: Results on VQA 2.0 test set.

4.5 Vision QA

We test on Visual Question Answering (VQA) 2.0 (Antol et al., 2015). Previous vision-language works, including those with language-generation functionality, almost all convert this task into a classification task: the models are trained to the answer from 3129 most frequent candidates (e.g., (Wang et al., 2022d)). We adopt a different open-vocabulary setting that i-Code V2 is trained to generate the answer. The task prompt is "Answer the following question based on the image:". Note that we don't provide candidate answer choices to i-Code V2 during testing. Table 5 contains baselines for both settings. i-Code V2's performance is competitive compared with vision-language models such as VisualBERT, LXMERT and VL-BERT. It is worth noting that Flamingo is pretrained on 2.1B vision-language data examples and has 80B parameters. In comparison i-Code V2 is pretrained on < 80M vision-language data and only has 1.4% parameters of Flamingo.

We then test i-Code V2 on VizWiz-VQA (Gurari et al., 2019), which is designed to answer visual questions from visually impaired people. Baselines include VisWiz Challenge Winner (Liu et al., 2021), BAN (Kim et al., 2018), B-Ultra & B-FRCNN (Changpinyo et al., 2019). i-Code V2 shows better performance than the previous VizWiz challenge winner and provides a strong baseline for models of intermediate size. As noted in Section 5, i-Code V2 also shows impressive zero-shot performance.

Model	Accuracy
BAN	51.6
B-FRCNN	51.9
B-Ultra	53.7
LXMERT	55.4
VisWiz Challenge Winner	60.6
Flamingo*	65.4
i-Code V2	**61.3**

Table 6: Performance on VisWiz-VQA test-std set.

Model	BLEU@4	METEOR	CIDEr	SPICE
VL-T5	34.5	28.7	116.5	21.9
VL-BART	-	-	116.6	-
BUTD	36.2	27.0	113.5	20.3
AoANet	**37.2**	28.4	119.8	21.3
UNITAB	35.8	28.4	119.1	21.5
XGPT	37.2	28.6	120.1	21.8
i-Code V2	36.8	**28.9**	**124.3**	**22.3**

Table 7: Experimental results (with cross-entropy optimization) on MS-COCO image captioning dataset (Karpathy test split).

4.6 Image Captioning

We evaluate i-Code V2 on the MS-COCO image captioning dataset (Chen et al., 2015) with the Karpathy test split (Karpathy and Fei-Fei, 2015), with results presented in Table 7. Evaluations metrics include BLEU@4, METEOR, CIDEr, and SPICE. The task prompt is "Generate the caption for this image: ". Baseline methods include image captioning models BUTD (Anderson et al., 2018) and AoANet (Huang et al., 2019); vision-language generative models, e.g., VL-BART, VL-T5 (Cho et al., 2021), XGPT(Xia et al., 2021); models using additional auxiliary input such as UNITAB (Yang et al., 2022) with object detection information. i-Code v2 outperforms vision-language baselines on METEOR, CIDEr, and SPICE.

5 Analysis & Explorations

Ablation study on pretraining effectiveness. We investigate the pretraining effectiveness by comparing performance of i-Code V2 with and without the multimodal pretraining (Section 3.2). As shown in Table 8, the multimodal pretraining further improves the performance on downstream tasks. The improvement is more significant on tasks where cross-modality understanding is more crucial, such as video sentiment analysis.

variant	UR-FUNNY	Image Chat		LibriSpeech
i-Code V2	Accuracy	F1	R-L	WER(%)↓
w/ pretraining	**79.59**	**15.5**	**18.6**	**3.86**
w/o pretraining	62.85	15.0	18.2	12.1

Table 8: Ablation study of the proposed multimodal pretraining.

Dataset	MSMO			ASR
Metric	R1	R2	RL	WER (↓)
w/o LGP	42.23	20.12	37.1	3.88
Full Pretraining	44.7	21.0	37.7	3.86

Table 9: Ablation study on pretraining objectives. LGP stands for "Language-only Generative Pretraining".

Pretraining objectives ablation. We explore how pretraining objectives affect model's performance. As shown in Table 9, removing "Language-only Generative Pretraining" (LGP) is adversarial for the performance on multimodal summarization, while it has negligible effect on ASR.

Zero-shot Learning. We test pretrained i-Code V2 on VizWiz-VQA without finetuning. As an open-vocabulary generative model, the zero-shot performance of i-Code V2 is respectful, with overall accuracy 22.53%, and 73.6% for "Yes/No" answers, 8.47% for "Number" answers, 24.46% for "Other" answers, 10.49% for "Unanswerable" answers respectively (for reference, Flamingo-9B zero-shot accuracy is 28.8%). Note that VizWiz-VQA questions are from visually impaired population and images are also distinct from those in VQA data used in pretraining. This performance indicates that i-Code V2 can closely follow the task instruction to answer the question. It also shows that i-Code V2 learns to answer visually grounded questions from pretraining, even though there are assorted pretraining tasks and datasets.

Training Hyperparameters. In Table 10, we list the learning rate, batch size (per GPU), and epochs for each finetuning dataset. We choose the finetuning checkpoint with the best performance on the validation test for the final evaluation. All finetuning jobs are conducted on eight A100 GPUs with AdamW optimizer.

6 Conclusion

In this paper, we propose i-Code V2, a multimodal generative model that jointly encodes language,

Task	lr	Batch Size	Epochs
MSMO	2.5×10^{-5}	8	12
Image Chat	2×10^{-5}	8	5
UR-FUNNY	1×10^{-5}	8	12
LibriSpeech	2×10^{-5}	2	10
VQA	2×10^{-6}	2	4
VisWiz-VQA	2×10^{-6}	2	4
MS-COCO	2×10^{-6}	2	4

Table 10: Training hyperparameters on downstream tasks.

vision and speech modalities and decodes the corresponding natural language sequence. i-Code V2 is pretrained on assorted high-quality single- and dual-modality datasets, where different tasks are unified as a multimodal sequence-to-sequence generation paradigm. i-Code V2 exhibits impressive performance in various multimodal generation domains, including multimodal nature language generation, ASR, vision QA, vision captioning and video sentiment analysis.

Limitations & Broader Impacts

i-Code V2 can inherit bias from the pretraining data, such as cultural or social bias. Since i-Code V2 has only been trained on English data, it is unclear how it extends to other languages. Fusing representations of different modalities other than concatenation is also a direction for future improvement. Including additional types of pretraining data, such as object detection, can help the model generalize to extra domains.

References

Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. 2021. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems*, 34.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. 2022. BEit: BERT pre-training of image transformers. In *International Conference on Learning Representations*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Soravit Changpinyo, Bo Pang, Piyush Sharma, and Radu Soricut. 2019. Decoupled box proposal and featurization with ultrafine-grained semantic labels improve image captioning and visual question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1468–1474.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX*, pages 104–120. Springer.

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, pages 1931–1942. PMLR.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Danna Gurari, Qing Li, Chi Lin, Yinan Zhao, Anhong Guo, Abigale Stangl, and Jeffrey P Bigham. 2019. Vizwiz-priv: A dataset for recognizing the presence and purpose of private visual information in images taken by blind people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 939–948.

Wei Han, Hui Chen, Alexander Gelbukh, Amir Zadeh, Louis-philippe Morency, and Soujanya Poria. 2021. Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, pages 6–15.

Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed Ehsan Hoque. 2019. Ur-funny: A multimodal language dataset for understanding humor. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2046–2056.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1122–1131.

Pengcheng He, Baolin Peng, Liyang Lu, Song Wang, Jie Mei, Yang Liu, Ruochen Xu, Hany Hassan Awadalla, Yu Shi, Chenguang Zhu, et al. 2022. Z-code++: A pre-trained language model optimized for abstractive summarization. *arXiv preprint arXiv:2208.09770*.

Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. 2019. Attention on attention for image captioning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4634–4643.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. 2023. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*.

Justin Johnson, Andrej Karpathy, and Li Fei-Fei. 2016. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4565–4574.

Da Ju, Kurt Shuster, Y-Lan Boureau, and Jason Weston. 2019. All-in-one image-grounded conversational agents. *arXiv preprint arXiv:1912.12394*.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137.

Jin-Hwa Kim, Yongseok Choi, Sungeun Hong, Jaehyun Jun, and Byoung-Tak Zhang. 2018. Bilinear attention networks for vizwiz challenge. In *Proceedings of the ECCV Workshop on VizWiz Grand challenge*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2020. What does bert with vision look at? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5265–5275.

Paul Pu Liang, Yiwei Lyu, Xiang Fan, Zetian Wu, Yun Cheng, Jason Wu, Leslie Yufan Chen, Peter Wu, Michelle A Lee, Yuke Zhu, et al. 2021. Multibench: Multiscale benchmarks for multimodal representation learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740.

Yu Liu, Lianghua Huang, Liuyihang Song, Bin Wang, Yingya Zhang, and Pan Pan. 2021. Enhancing textual cues in multi-modal transformers for vqa. In *2021 VizWiz Grand Challenge Workshop*.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. 2018. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th Annual Meeting of* *the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2247–2256.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. 2022. Unified-io: A unified model for vision, language, and multi-modal tasks. *arXiv preprint arXiv:2206.08916*.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2630–2640.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

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

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M Smith, et al. 2020. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo-Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. 2022. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Suwon Shon, Ankita Pasad, Felix Wu, Pablo Brusco, Yoav Artzi, Karen Livescu, and Kyu J Han. 2022. Slue: New benchmark tasks for spoken language understanding evaluation on natural speech. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7927–7931. IEEE.

Kurt Shuster, Samuel Humeau, Antoine Bordes, and Jason Weston. 2020. Image-chat: Engaging grounded conversations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2414–2429.

Kurt Shuster, Eric Michael Smith, Da Ju, and Jason Weston. 2021. Multi-modal open-domain dialogue. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4863–4883.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing* *and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111.

Zineng Tang, Ziyi Yang, Guoxin Wang, Yuwei Fang, Yang Liu, Chenguang Zhu, Michael Zeng, Cha Zhang, and Mohit Bansal. 2022. Unifying vision, text, and layout for universal document processing. *arXiv preprint arXiv:2212.02623*.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569.

Changhan Wang, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Ann Lee, Peng-Jen Chen, Jiatao Gu, and Juan Pino. 2021. fairseq s^2: A scalable and integrable speech synthesis toolkit. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 143–152.

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022a. GIT: A generative image-to-text transformer for vision and language. *Transactions on Machine Learning Research*.

Junke Wang, Dongdong Chen, Zuxuan Wu, Chong Luo, Luowei Zhou, Yucheng Zhao, Yujia Xie, Ce Liu, Yu-Gang Jiang, and Lu Yuan. 2022b. OmniVL: One foundation model for image-language and video-language tasks. In *Advances in Neural Information Processing Systems*.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022c. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR.

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

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2022e. SimVLM: Simple visual language model pretraining with weak supervision. In *International Conference on Learning Representations*.

Qiaolin Xia, Haoyang Huang, Nan Duan, Dongdong Zhang, Lei Ji, Zhifang Sui, Edward Cui, Taroon Bharti, and Ming Zhou. 2021. Xgpt: Cross-modal generative pre-training for image captioning. In *Natural Language Processing and Chinese Computing: 10th CCF International Conference, NLPCC 2021, Qingdao, China, October 13–17, 2021, Proceedings, Part I 10*, pages 786–797. Springer.

Xiao Xu, Chenfei Wu, Shachar Rosenman, Vasudev Lal, and Nan Duan. 2022. Bridge-tower: Building bridges between encoders in vision-language representation learning. *arXiv preprint arXiv:2206.08657*.

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Faisal Ahmed, Zicheng Liu, Yumao Lu, and Lijuan Wang. 2022. Unitab: Unifying text and box outputs for grounded vision-language modeling. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVI*, pages 521–539. Springer.

Ziyi Yang, Yuwei Fang, Chenguang Zhu, Reid Pryzant, Dongdong Chen, Yu Shi, Yichong Xu, Yao Qian, Mei Gao, Yi-Ling Chen, et al. 2023. i-code: An integrative and composable multimodal learning framework. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Dong Yu and Li Deng. 2016. *Automatic speech recognition*, volume 1. Springer.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*, Aug 2022.

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

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246.

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

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

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.

Zhengkun Zhang, Xiaojun Meng, Yasheng Wang, Xin Jiang, Qun Liu, and Zhenglu Yang. 2022. Unims: A unified framework for multimodal summarization with knowledge distillation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11757–11764.

Junnan Zhu, Haoran Li, Tianshang Liu, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2018. Msmo: Multimodal summarization with multimodal output. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 4154–4164.

Junnan Zhu, Yu Zhou, Jiajun Zhang, Haoran Li, Chengqing Zong, and Changliang Li. 2020. Multimodal summarization with guidance of multimodal reference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9749–9756.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.