LAVIS: A One-stop Library for Language-Vision Intelligence

Dongxu Li, Junnan Li, Hung Le, Guangsen Wang, Silvio Savarese, Steven C.H. Hoi
Salesforce Research
Open-source repository: https://github.com/salesforce/LAVIS
Supplementary video: https://youtu.be/0CuRowHu7TA

Abstract

We introduce LAVIS, an open-source deep learning library for LAnguage-VISion research and applications. LAVIS aims to serve as a one-stop comprehensive library that brings recent advancements in the language-vision field accessible for researchers and practitioners, as well as fertilizing future research and development. It features a unified interface to easily access state-of-the-art image-language, video-language models and common datasets. LAVIS supports training, evaluation and benchmarking on a rich variety of tasks, including multimodal classification, retrieval, captioning, visual question answering, dialogue and pre-training. In the meantime, the library is also highly extensible and configurable, facilitating future development and customization. In this paper, we describe design principles, key components and functionalities of the library, and also present benchmarking results across common language-vision tasks.

1 Introduction

Multimodal content, in particular language-vision data including texts, images and videos are ubiquitous for real-world applications, such as content recommendation, e-commerce and entertainment. There has been tremendous recent progress in developing powerful language-vision models (Su et al., 2020; Lu et al., 2019; Chen et al., 2020; Li et al., 2020; Huang et al., 2021; Li et al., 2021a; Radford et al., 2021; Zhou et al., 2020; Gan et al., 2020; Cho et al., 2021; Zhang et al., 2021; Li et al., 2022b; Zhu and Yang, 2020; Bain et al., 2021; Xu et al., 2021; Lei et al., 2021; Li et al., 2022a). However, training and evaluating these models across tasks and datasets require domain knowledge and are not always welcoming to incoming researchers and practitioners. This is mainly due to inconsistent interfaces across models, datasets and task evaluations, and also the duplicating yet non-trivial efforts to prepare the required experiment setup. To make accessible the emerging language-vision intelligence and capabilities to a wider audience, promote their practical adoptions, and reduce repetitive efforts in future development, we build LAVIS (short for LAnguage-VISion), an open-source library for training, evaluating state-of-the-art language-vision models on a rich family of common tasks and datasets, as well as for off-the-shelf inference on customized language-vision data.

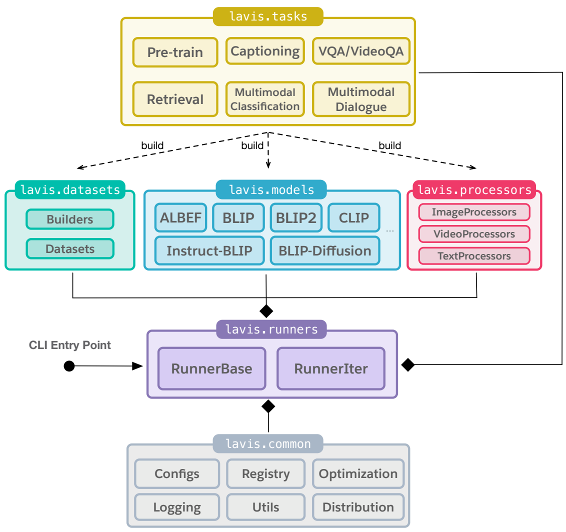

Figure 1: Overall architecture of the LAVIS library.

Figure 1 shows the overall design of LAVIS. Important features of LAVIS include (i) **Unified interface and modular design**. Key components in the library are organized using a unified and modular design. This allows effortless off-the-shelf access to individual components, swift development and easy integration of new or external components. The modular design also eases model inferences, such as multimodal feature extraction. (ii) **Comprehensive support of image-text, video-**

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 3: System Demonstrations, pages 31–41
July 10-12, 2023 ©2023 Association for Computational Linguistics

text tasks and datasets. LAVIS supports a growing list of more than ten common language-vision tasks, across over 20 public datasets. These tasks and datasets provide a comprehensive and unified benchmark for evaluating language-vision models. (iii) **State-of-the-art and reproducible language-vision models**. The library enables access to over 30 pre-trained and task-specific fine-tuned model checkpoints of 5 foundation models: ALBEF (Li et al., 2021a), BLIP (Li et al., 2022b), BLIP2 (Li et al., 2023b), CLIP (Radford et al., 2021) and ALPRO (Li et al., 2022a), as well as state-of-the-art language-vision methods such as PnP-VQA (Tiong et al., 2022), Img2Prompt (Guo et al., 2022). These models achieve competitive performance across multiple tasks, representing the up-to-date development status of the language-vision research. We also provide training, evaluation scripts and configurations to facilitate reproducible language-vision research and adoption. (iv) **Resourceful and useful toolkit**. In addition to the core library functionalities, we also provide useful resources to reduce the learning barriers for the language-vision research. This includes automatic dataset downloading tools to help prepare the supported datasets, a GUI dataset browser to help preview downloaded datasets and dataset cards documenting sources, supported tasks and leaderboards.

2 Related Work

Table 1 summarizes the comparisons between LAVIS' key features with those of other libraries. Most related libraries include MMF (Singh et al., 2020), UniLM (uni, 2020), X-modaler (Li et al., 2021b) and TorchMultimodal (tor, 2022).

- MMF is a comprehensive multimodal framework encapsulating many language-vision models and datasets. It implements modular interface for training and evaluation. However, it consists of mostly task-specific architectures. Besides showing relatively inferior performance, these models are usually not easy to transfer across tasks. Among the included foundation models (Li et al., 2019; Chen et al., 2020; Zhang et al., 2021; Li et al., 2021a) in MMF, few fully supports finetuning or benchmarking on the extended list of downstream tasks. In contrast, considering that pre-trained foundation models prevail across overwhelmingly many tasks and datasets with more principal and unified architectures, our library focuses on pre-trained models and their task-specific variants instead.
- UniLM was initiated for developing large language models, and recently also aggregates multiple standalone repositories of multimodal models. Yet, support for multimodal models in UniLM is limited in its current development status. Moreover, UniLM does not provide unified or modular interfaces to allow easy access or reproduction.
- X-modaler supports a limited number of tasks and datasets, which are not as comprehensive as LAVIS. Besides, similar to MMF, models in X-modaler are also mostly in task-specific architectures. The few supported foundation model, e.g. (Chen et al., 2020), achieves inferior results than models in LAVIS.
- A concurrent library TorchMultimodal (tor, 2022) promotes modular development of language-vision models. Our library supports a wider range of tasks and datasets than TorchMultimodal while being more comprehensive and resourceful.

Other open-source implementations of individual models exist (Chen et al., 2020; Li et al., 2020; Lu et al., 2019; Radford et al., 2021; Gan et al., 2020; Lei et al., 2021), yet do not provide centralized access. In summary, in contrast to previous efforts, our library stands out by providing *easier* access to *stronger* models on comprehensively *many* tasks and datasets. With this effort, we hope to significantly reduce the cost and effort to leverage and benchmark existing multimodal models, as well as to develop new models.

3 Supported Tasks, Datasets and Models

Table 3 summarizes the supported tasks, datasets and models in LAVIS. In particular, we prioritize tasks that are standard, widely adopted for evaluation, and with publicly available datasets. For image-text tasks, the library implements image-text retrieval, image captioning, visual question answering (VQA), visual dialogue, visual entailment (VE), natural language visual reasoning ($NLVR^2$) and image classification. For video-text tasks, LAVIS currently support video-text retrieval and video question answering (VideoQA). There are in total over 20 public datasets supported, including MSCOCO (Lin et al., 2014), Flickr30k (Plummer et al., 2015), VQAv2 (Goyal et al., 2017), OK-VQA (Marino

Table 1: Comparison of features in LAVIS and other existing language-vision libraries or codebase. Note that language-vision models in UniLM and TorchMultimodal (alpha release) are under development, therefore, the table only includes their supported features by the publication time of this technical report.

		LAVIS (Ours)	MMF	UniLM	X-modaler	TorchMultimodal
Unified Model and Dataset Interface		✓				
Modular Library Design		✓	✓		✓	✓
Pre-trained Model Checkpoints		✓				
Task-specific Finetuned Model Checkpoints		✓			✓	
Modalities	Image-Text	✓	✓	✓	✓	✓
	Video-Text	✓	✓		✓	
Tasks	End2end Pre-training	✓		✓		✓
	Multimodal Retrieval	✓	✓		✓	
	Captioning	✓	✓		✓	
	Visual Question Answering	✓	✓		✓	
	Multimodal Classification	✓	✓			
	Instructed Zero-shot Generation	✓				
	Visual Dialogue	✓				
	Multimodal Feature Extraction	✓				
Toolkit	Benchmarks	✓				
	Dataset Auto-downloading	✓	✓			
	Dataset Browser	✓				
	GUI Demo	✓				
	Dataset Cards	✓				

Table 2: Supported tasks, datasets and models in LAVIS.

Supported Tasks	**Supported Models**	**Supported Datasets**
Image-text Pre-training	ALBEF, BLIP, BLIP2, InstructBLIP	COCO, Visual Genome, SBU Caption, Conceptual Captions (3M, 12M), LAION
Image-text Retrieval	ALBEF, BLIP, BLIP2, CLIP	COCO, Flickr30k
Visual Question Answering	ALBEF, BLIP, BLIP2, InstructBLIP	VQAv2, OKVQA, A-OKVQA, GQA
Image Captioning	BLIP, BLIP2, InstructBLIP	COCO Caption, NoCaps
Image Classification	CLIP	ImageNet
Natural Language Visual Reasoning (NLVR^2)	ALBEF, BLIP	NLVR^2
Visual Entailment	ALBEF	SNLI-VE
Visual Dialogue	BLIP, InstructBLIP	VisDial
Video-text Retrieval	ALPRO, BLIP	MSRVTT, DiDeMo
Video Question Answering	ALPRO, BLIP, InstructBLIP	MSRVTT-QA, MSVD-QA
Video Dialogue	BLIP	AVSD

et al., 2019), A-OK-VQA (Shevchenko et al., 2021), GQA (Hudson and Manning, 2019), Visual Genome (Krishna et al., 2017), ImageNet (Deng et al., 2009), NoCaps (Agrawal et al., 2019), Conceptual Captions (Sharma et al., 2018; Changpinyo et al., 2021), SBU-caption (Ordonez et al., 2011), LAION (Schuhmann et al., 2021), NLVR^2 (Suhr et al., 2019), SNLI-VE (Bowman et al., 2015), VisDial (Das et al., 2017), AVSD (Alamri et al., 2019), MSRVTT (Xu et al., 2016), MSVD (Xu et al., 2017), DiDeMo (Anne Hendricks et al., 2017) and their task-specific variants. LAVIS currently supports 6 foundation models, i.e. ALBEF (Li et al., 2021a), BLIP (Li et al., 2022b), BLIP2 (Li et al., 2023b), CLIP (Radford et al., 2021), InstructBLIP (Dai et al., 2023) and ALPRO (Li et al., 2022a). In addition, the library also features language-vision methods including PnP-VQA (Tiong et al., 2022) and Img2prompt (Guo et al., 2022), and text-to-image generation model BLIP-Diffusion (Li et al., 2023a). These models and methods show strong performance on the aforementioned tasks and datasets, representing the up-to-date development status of the language-vision research field. Detailed description can be found in A.1

4 Library Design

This section delineates the design of LAVIS as shown in Figure 1. Our key design principle is to provide a simple and unified library to easily (i) train and evaluate the model; (ii) access supported models and datasets; (iii) extend with new models, tasks and datasets.

4.1 Description on each library component

Key components in LAVIS include:

- **Runners** – `lavis.runners` module manages the overall training and evaluation lifecycle. It is also responsible for creating required components lazily as per demand, such as optimizers, learning rate schedulers and dataloaders. Currently, `RunnerBase` implements epoch-based training and `RunnerIters` implements iteration-based training.

- **Tasks** – `lavis.tasks` module implements concrete training and evaluation logic per task. This includes pre-training and finetuning tasks as listed in Table 3. The rationale to have an abstraction of task is to accommodate task-specific training, inference and evaluation. For example, evaluating a retrieval model is different from a classification model.

- **Datasets** – `lavis.datasets` module helps create datasets. Specifically, `datasets.builders` module loads dataset configurations, downloads annotations and builds the dataset;

  - `lavis.datasets.datasets` module defines the supported datasets, each is a PyTorch dataset instance.
  - We also provide automatic dataset downloading tools in `datasets/download_scripts` to help prepare common public datasets.

- **Models** – `lavis.models` module holds definitions for the supported models and shared model layers.

- **Processors** –`lavis.processors` module handles preprocessing of multimodal input. A processor transforms input images, videos and texts into the desired form that models can consume.

- **Common tools and utilities** – `lavis.commons` module contains shared classes and methods used by multiple other modules. For example, `configs` module contains classes to store and manipulate configuration files used by LAVIS. In particular, we use a hierarchical configuration design, to allow highly customizable training and evaluation. The `registry` module serves as a centralized place to manage modules that share the same functionalities. It allows building datasets, models, tasks, and learning rate schedulers during runtime, by specifying their names in the configuration; `optims` contains definitions of learning rate schedulers; `utils` contains miscellaneous utilities, mostly IO-related helper functions;

4.2 Example library usage

The design of the library enables easy access to existing models and future development. In this section, we include a few examples to demonstrate some common use cases.

Unified interface for data and model loading

LAVIS provides unified interface `load_dataset` and `load_model` to access supported datasets and models. This is helpful for off-the-shelf use of datasets and model inference etc. In the first example, we show how to load a dataset using the library.

```
from lavis.datasets.builders import
    load_dataset
# load a specific dataset
coco_dataset = load_dataset("
    coco_caption")
# dataset is organized by split names.
print(coco_dataset.keys())
# dict_keys(['train', 'val', 'test'])
# total number of samples in the
    training split.
print(len(coco_dataset["train"]))
# 566747
# peek a random sample
print(coco_dataset["train"][0])
# {'image': <PIL.Image.Image image mode=
    RGB size=640x480>,
#   'text_input': 'A woman wearing a net
     on her head cutting a cake. ',
#   'image_id': 0}
```

Models and their related preprocessors can also be loaded via a unified interface, which facilitates effortless analysis and inference on custom data. In the following, we show an example that uses a BLIP captioning model to generate image captions.

```
from lavis.models import
    load_model_and_preprocess
# load model and preprocessors
model, vis_procs, _ =
    load_model_and_preprocess(
  name="blip_caption", model_type="
    base_coco")
```

```python
# raw_image is a PIL Image instance
raw_image = coco_dataset["test"][0]["image"]
# preprocess a raw input image
image = vis_procs["eval"](raw_image).unsqueeze(0)
# generate caption
caption = model.generate({"image": image})
# ['a man riding a motorcycle down a dirt road']
```

Unified interface for multimodal feature extraction

LAVIS supports a unified interface to extract multimodal features. The features are useful especially for offline applications where end-to-end finetuning is not affordable. By changing name and model_type, users can choose to use different model architecture and pre-trained weights.

```python
# load feature extraction models and processors
model, vis_procs, txt_procs = load_model_and_preprocess(
        name="blip_feature_extractor",
        model_type="base"
)
# a random instance from coco dataset
raw_image = coco_dataset["test"][0]["image"]
text = coco_dataset["test"][0]["text_input"]
# process the input
image = vis_procs["eval"](raw_image).unsqueeze(0)
text_input = txt_procs["eval"](text)
sample = {"image": image,
          "text_input": [text_input]}

# extract multimodal features
feature = model.extract_features(sample)
```

5 Benchmarks and Library Toolkit

In this section, we benchmark model performance across tasks and datasets in LAVIS. Then we take our web demo interface to show a few case studies on multimodal content understanding. We also present a GUI dataset browser that helps to preview supported datasets.

5.1 Main results

The purpose of the benchmark is two-fold. First, we use the benchmark to validate that our re-implementation faithfully replicates official models. Second, the benchmark also serves as a reference for further development. In Table 5-4, we organize benchmark results by models and compare our replication results with those reported officially. Experiments are conducted on NVIDIA A100 GPUs.

Table 3: Comparison between official and replicated performance using BLIP. TR denotes text retrieval; IR denotes image retrieval. Results are produced by $\text{BLIP}_{\text{CapFilt-L}}$ model. NoCaps results are reported on the entire validation set. Retrieval and captioning results are reported on the test sets; B@4 denotes BLEU-4.

Tasks	Datasets	Impl.	Results		
Retrieval			**R1**	**R5**	**R10**
TR	COCO		82.4	95.4	97.9
			82.0	95.8	98.1
IR	COCO		65.1	86.3	91.8
			64.5	86.0	91.7
TR	Flickr30k		97.2	99.9	100.0
			96.9	99.9	100.0
IR	Flickr30k		87.5	97.7	98.9
			87.5	97.6	98.9
			dev	**std**	
VQA	VQAv2		78.25	78.32	
			78.23	78.29	
			B@4	**CIDEr**	**SPICE**
Image Captioning	COCO		39.7	133.3	-
			39.7	133.5	23.7
	NoCaps		-	109.6	14.7
			31.9	109.1	14.7
			val	**test**	
Multimodal Classification	NLVR2		82.15	82.24	
			82.48	83.25	

Table 4: Comparison between official and replicated performance using CLIP-ViT-L/336. Note the relative difference is possibly due to the versioning of the model weights.

Tasks	Datasets	Impl.	Results		
Retrieval			**R1**	**R5**	**R10**
TR	COCO		58.4	81.5	88.1
			57.2	80.5	87.8
IR	COCO		37.8	62.4	72.2
			36.5	60.8	71.0
TR	Flickr30k		88.0	98.7	99.4
			86.5	98.0	99.1
IR	Flickr30k		68.7	90.6	95.2
			67.0	88.9	93.3
			val		
Zero-shot Image Classification	ImageNet		76.2		
			76.5		

For ALBEF, BLIP, BLIP2 and ALPRO, we re-implement their models in LAVIS based on the official repositories and report finetuning results using their official pre-trained weights. For CLIP models, we integrate a third-party implementation (Ilharco et al., 2021) and report CLIP-ViT-L/336 zero-shot inference results using the official weights (Radford et al., 2021) (Table 4). As can be seen in the

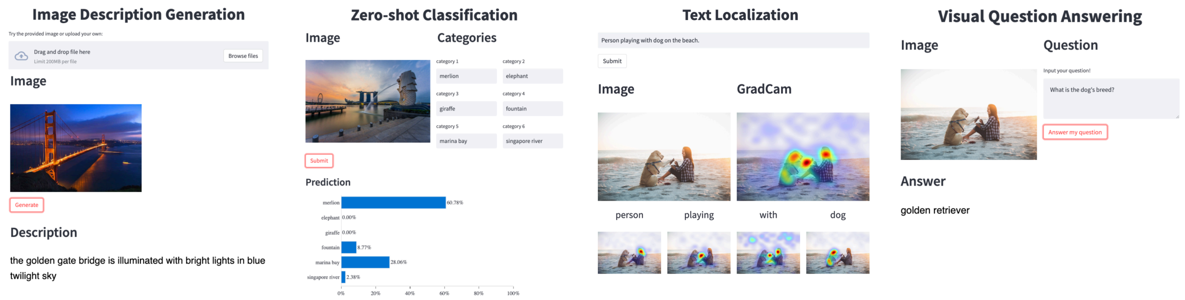

Figure 2: Screenshots of the GUI web demo, showing various applications including image captioning, zeros-shot image classification, text localization and visual question answering.

Figure 3: The developed dataset browser helps to quickly gain understanding of multimodal datasets.

tables, our library produce consistent results as reported officially. More benchmarking results with BLIP, ALPRO models can be found in A.2.

5.2 Library resources and toolkit

In addition to the components aforementioned, LAVIS also provides useful toolkit and resources to further ease development. This includes pre-trained and finetuned model checkpoints, automatic dataset downloading tools, a web demo and a dataset browser.

Pre-trained and finetuned model checkpoints. We include pre-trained and finetuned model checkpoints in the library. This promotes easy replication of our experiment results and to repurpose pre-trained models for other applications. Model checkpoints are downloaded automatically upon loading models.

Web demo. As shown in Figure 2, we develop a GUI-based web demo, which aims to provide a user-friendly interface to explore various multimodal capabilities. Currently the demo supports the following functionalities: (i) *image captioning*: produces a caption in natural language to describe an input image; (ii) *visual question answering*: answer natural language questions regarding the input image; (iii) *multimodal search*: search images in a gallery given a text query; (iv) *text visualization*: given an input image and a text caption, produces GradCam (Selvaraju et al., 2017) for each text token on the image; (v) zero-shot multimocal classification: classify an input images into a set of input labels in text. (vi) Thanks to the modular design of LAVIS, one can easily extend the demo with new functionalities, such as *text-to-image generation*, as shown in the Figure 2.

Automatic dataset downloading and browsing. Preparing language-vision datasets for pre-training and fine-tuning incurs much duplicating effort. To this end, LAVIS provides tools to automatically download and organize the public datasets, so that users can get access to the common datasets easier and quicker. In addition, we develop a GUI dataset browser, as shown in Figure 3, that helps users to rapidly gain intuitions about the data they use.

6 Conclusion and Future Work

We present LAVIS, an open-source deep learning library for language-vision research and applications. The library is designed to provide researchers and practitioners with easier and comprehensive access to state-of-the-art multimodal capabilities, The library also features a unified interface and extensible design to promote future development. Besides, the library also features extensive access to pre-trained weights and useful resources to reduce duplicating replication efforts. With these features, we expect LAVIS to serve as a one-stop library in multimodal AI for a wider audience.

We continue to actively develop and improve LAVIS. In future releases, our priorities are to include more language-vision models, tasks and datasets to the library. We also plan to add more parallelism support for scalable training and inference. While we will maintain LAVIS in the long term, we invite contributions from the open-source community to join this evolving effort.

Broader Impact and Responsible Use

LAVIS can provide useful capabilities for many real-world multimodal applications. It features easy, unified and centralized access to powerful language-vision models, facilitating effective multimodal analysis and reproducible research and development. We encourage researchers, data scientists, and ML practitioners to adopt LAVIS in real-world applications for positive social impacts, e.g. efficient and environment-friendly large-scale multimodal analysis.

However, LAVIS may also be misused. We encourage users to read detailed discussion and guidelines for building responsible AI, e.g. (Baxter, 2022). In particular, LAVIS should not be used to develop multimodal models that may expose unethical capabilities.

It is also important to note that that models in LAVIS provide no guarantees on their multimodal abilities; incorrect or biased predictions with out-of-date information may be observed. In particular, the datasets and pretrained models utilized in LAVIS contain socioeconomic biases which may result in misclassification and other unwanted behaviors such as offensive or inappropriate speech. We strongly recommend that users review the pretrained models and overall system in LAVIS before practical adoption. We plan to improve the library by investigating and mitigating these potential biases and inappropriate behaviors in the future.

References

2020. Large-scale self-supervised pre-training across tasks, languages, and modalities. https://github.com/microsoft/unilm.

2022. Torchmultimodal (alpha release). https://github.com/facebookresearch/multimodal.

Harsh Agrawal, Peter Anderson, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. nocaps: novel object captioning at scale. In *ICCV*, pages 8947–8956.

Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K Marks, Chiori Hori, Peter Anderson, et al. 2019. Audio visual scene-aware dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7558–7567.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812.

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 *ICCV*.

Kathy Baxter. 2022. Ai is everywhere — but are you building it responsibly? https://www.salesforce.com/blog/build-ethical-ai/?hasLoggedIn=true.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 2021. Is space-time attention all you need for video understanding? In *ICML*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *EMNLP*, pages 632–642.

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

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 *ECCV*, volume 12375, pages 104–120.

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. *arXiv preprint arXiv:2102.02779*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *CVPR*, pages 1080–1089.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186.

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. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*.

Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. 2020. Large-scale adversarial training for vision-and-language representation learning. In *NeurIPS*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *CVPR*, pages 6325–6334.

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

Jiaxian Guo, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Boyang Li, Dacheng Tao, and Steven CH Hoi. 2022. From images to textual prompts: Zero-shot vqa with frozen large language models. *arXiv preprint arXiv:2212.10846*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. 2021. Seeing out of the box: End-to-end pre-training for vision-language representation learning. *arXiv preprint arXiv:2104.03135*.

Drew A. Hudson and Christopher D. Manning. 2019. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, pages 6700–6709.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. Openclip. If you use this software, please cite it as below.

Amita Kamath, Christopher Clark, Tanmay Gupta, Eric Kolve, Derek Hoiem, and Aniruddha Kembhavi. 2022. Webly supervised concept expansion for general purpose vision models. *arXiv preprint arXiv:2202.02317*.

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 Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123(1):32–73.

Hung Le, Nancy Chen, and Steven Hoi. 2022. VGNMN: Video-grounded neural module networks for video-grounded dialogue systems. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3377–3393, Seattle, United States. Association for Computational Linguistics.

Hung Le, Nancy F. Chen, and Steven Hoi. 2021. Learning reasoning paths over semantic graphs for video-grounded dialogues. In *International Conference on Learning Representations*.

Hung Le and Steven C.H. Hoi. 2020. Video-grounded dialogues with pretrained generation language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5842–5848, Online. Association for Computational Linguistics.

Hung Le, Doyen Sahoo, Nancy Chen, and Steven Hoi. 2019. Multimodal transformer networks for end-to-end video-grounded dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5612–5623, Florence, Italy. Association for Computational Linguistics.

Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. 2021. Less is more: Clipbert for video-and-language learning via sparse sampling. In *CVPR*, pages 7331–7341.

Dongxu Li, Junnan Li, and Steven C. H. Hoi. 2023a. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing.

Dongxu Li, Junnan Li, Hongdong Li, Juan Carlos Niebles, and Steven C.H. Hoi. 2022a. Align and prompt: Video-and-language pre-training with entity prompts. In *CVPR*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022b. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*.

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. 2021a. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*.

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*, abs/1908.03557.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, pages 121–137.

Yehao Li, Yingwei Pan, Jingwen Chen, Ting Yao, and Tao Mei. 2021b. X-modaler: A versatile and high-performance codebase for cross-modal analytics. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3799–3802.

Zekang Li, Zongjia Li, Jinchao Zhang, Yang Feng, and Jie Zhou. 2021c. Bridging text and video: A universal multimodal transformer for audio-visual scene-aware dialog. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 29:2476–2483.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *ECCV*, volume 8693, pages 740–755.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, pages 13–23.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204.

Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. 2011. Im2text: Describing images using 1 million captioned photographs. In *NIPS*, pages 1143–1151.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, pages 2641–2649.

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. *arXiv preprint arXiv:2103.00020*.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, pages 2556–2565.

Violetta Shevchenko, Damien Teney, Anthony Dick, and Anton van den Hengel. 2021. Reasoning over vision and language: Exploring the benefits of supplemental knowledge. *arXiv preprint arXiv:2101.06013*.

Amanpreet Singh, Vedanuj Goswami, Vivek Natarajan, Yu Jiang, Xinlei Chen, Meet Shah, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. 2020. Mmf: A multimodal framework for vision and language research. `https://github.com/facebookresearch/mmf`.

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 *ICLR*.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. A corpus for reasoning about natural language grounded in photographs. In *ACL*, pages 6418–6428.

Anthony Meng Huat Tiong, Junnan Li, Boyang Li, Silvio Savarese, and Steven CH Hoi. 2022. Plug-and-play vqa: Zero-shot vqa by conjoining large pre-trained models with zero training. *arXiv preprint arXiv:2210.08773*.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. 2017. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the ACM international conference on Multimedia*, pages 1645–1653.

Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. 2021. Videoclip: Contrastive pre-training for zero-shot video-text understanding. In *EMNLP*, pages 6787–6800.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: Making visual representations matter in vision-language models. *arXiv preprint arXiv:2101.00529*.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. 2020. Unified vision-language pre-training for image captioning and VQA. In *AAAI*, pages 13041–13049.

Linchao Zhu and Yi Yang. 2020. Actbert: Learning global-local video-text representations. In *CVPR*, pages 8746–8755.

A Appendix

A.1 Details of the supported models

- ALBEF is an image-text model. It employs a ViT (Dosovitskiy et al., 2021) as the image encoder, early BERT (Devlin et al., 2019) layers as the text encoder, and re-purposes late BERT layers as the multimodal encoder by adding cross-attentions. It proposes the novel image-text contrastive (ITC) loss to align unimodal features before fusing them using the multimodal encoder. It is also one of the first few models requiring no region information while demonstrating strong multimodal understanding capability.

- BLIP primarily tackles image-text tasks, while also showing strong zero-shot transfer capabilities to video-text tasks. It employs a ViT as the image encoder and a BERT as the text encoder. To facilitate multimodal understanding and generation, BLIP proposes mixture of encoder-decoder (MED), which re-purposes BERT into multimodal encoder and decoder with careful weight sharing. Moreover, BLIP proposes dataset bootstrapping to improve the quality of texts in the pre-training corpus by removing noisy ones and generating new diverse ones. In addition to the improved understanding capability compared to ALBEF, BLIP highlights its strong text generation ability, producing accurate and descriptive image captions. When adapted to video-text tasks, it operates on sampled frames while concatenating their features to represent the video.

- BLIP2 represents a generic and efficient language-vision pre-training strategy that leverages available frozen image encoders and large language models (LLMs). The model introduces a two-staged training strategy to bridge the modality gap with a lightweight module, called Querying Transformer (Q-Former). In addition to the strong performance on existing tasks, including VQA, multimodal retrieval, captioning, BLIP-2 also unlocks the novel capabilities of zero-shot image-to-text generation following natural language instructions.

- CLIP is a family of powerful image-text models. Different from ALBEF and BLIP, CLIP models adopt two unimodal encoders to obtain image and text representations. CLIP maximizes the similarity between positive image-text pairs, and was trained on 400M image-text pairs, rendering strong and robust unimodal representations. CLIP variants employ different visual backbones, including ResNet-50 (He et al., 2016), ViT-B/16, ViT-B/32, ViT-L/14, ViT-L/14-336. We integrate a third-party implementation of CLIP (Ilharco et al., 2021) into LAVIS while including the official pre-trained weights.

- ALPRO is a video-text model, tackling video-text retrieval and video question answering tasks. It uses TimeSformer (Bertasius et al., 2021) to extract video features, and BERT to extract text features. Similar to ALBEF, ALPRO uses contrastive loss to align unimodal features, yet it opts to use self-attention to model multimodal interaction. This architecture choice enables an additional visual-grounded pre-training task, i.e. prompt entity modeling (PEM) to align fine-grained video-text information. ALPRO is strong in extracting regional video features and remains competitive for video understanding tasks across various datasets.

A.2 Additional benchmarking results

In Table 5 and 6, we show benchmarking results with BLIP and ALPRO reimplmentations in LAVIS. As shown in the tables, the results are consistent with those in the original implementation.

In Table 7, we present results by adapting models in LAVIS to new tasks and datasets, on which the models were not previously reported on. In this way, we show that our library helps to easily adapt to new tasks and datasets, while achieving competitive performance.

Knowledge-based VQA (KVQA). The task of KVQA aims to measure the commonsense knowledge learnt by language-vision models, where models are asked to answer questions involving external knowledge. To this end, state-of-the-art models (Gui et al., 2021; Kamath et al., 2022) resort to external knowledge base(Vrandečić and Krötzsch, 2014) or large language models(Brown et al., 2020). In our experiments, we show that language-vision pre-trained models finetuned on VQAv2(Goyal et al., 2017) show strong transfer results to KVQA datasets. With additional finetuning on KVQA datasets, further improvements are ob-

Table 5: Comparison between official and replicated performance using BLIP. TR denotes text retrieval; IR denotes image retrieval. Results are produced by $\text{BLIP}_{\text{CapFilt-L}}$ model. NoCaps results are reported on the entire validation set. Retrieval and captioning results are reported on the test sets; B@4 denotes BLEU-4.

Tasks	Datasets	Impl.	Results		
Retrieval			**R1**	**R5**	**R10**
TR	COCO		82.4	95.4	97.9
			82.0	95.8	98.1
IR	COCO		65.1	86.3	91.8
			64.5	86.0	91.7
TR	Flickr30k		97.2	99.9	100.0
			96.9	99.9	100.0
IR	Flickr30k		87.5	97.7	98.9
			87.5	97.6	98.9
			dev	**std**	
VQA	VQAv2		78.25	78.32	
			78.23	78.29	
			B@4	**CIDEr**	**SPICE**
Image Captioning	COCO		39.7	133.3	-
			39.7	133.5	23.7
	NoCaps		-	109.6	14.7
			31.9	109.1	14.7
			val	**test**	
Multimodal Classification	NLVR2		82.15	82.24	
			82.48	83.25	

served on both OK-VQA and AOK-VQA datasets. As a result, our best model BLIP surpasses previous state-of-the-art by a clear margin.

Video Dialogue. The task of video-grounded dialogues requires models to generate a natural response given a dialogue context and a grounding video (Alamri et al., 2019). Existing models have exploited new architectural designs (Le et al., 2019), additional learning tasks (Le et al., 2022, 2021), and pretraining (Le and Hoi, 2020; Li et al., 2021c) to improve the model abilities to understand multimodal context and generate natural language. In our experiments, we show that our library can be easily integrated with any vision-language models (such as VGD-GPT (Le and Hoi, 2020)) to adapt to this dialogue task. The results in Table 7 show that our model implementation with LAVIS can lead to impressive performance, comparable to current state-of-the-art approaches.

A.3 Supplementary video and online demo:

The supplementary video can be found:`https://youtu.be/0CuRowHu7TA`. In the following, we provide additional benchmarking results using models in LAVIS.

Alternatively, the video can be downloaded from: `https://drive.google.com/file/d/1cFTEgL53WI-oSFbWR_6k6eg2iqAi9bwe/view?usp=sharing`

Table 6: Comparison between official and replicated task performance using ALPRO. TR denotes video-to-text retrieval; VR denotes text-to-video retrieval.

Tasks	Datasets	Impl.	Results		
Retrieval			**R1**	**R5**	**R10**
TR	MSRVTT		32.0	60.7	70.8
			33.2	60.5	71.7
VR	MSRVTT		33.9	60.7	73.2
			33.8	61.4	72.7
TR	DiDeMo		37.9	67.1	77.9
			38.8	66.4	76.8
VR	DiDeMo		35.9	67.5	78.8
			36.6	67.5	77.9
			test		
VideoQA	MSRVTT		42.1		
			42.1		
	MSVD		45.9		
			46.0		

Table 7: Experiment results on KVQA compared with best existing methods. Due to the submission number limits, only BLIP AOKVQA result on the test split is reported.

Tasks	Datasets	Models	Results	
			test	
KVQA	OKVQA	KAT (Single)(Gui et al., 2021)	53.1	
		KAT (Ensemble)(Gui et al., 2021)	54.4	
		ALBEF	54.7	
		BLIP	55.4	
			val	**test**
	AOKVQA	GPV-2(Kamath et al., 2022)	48.6	40.7
		ALBEF	54.5	-
		BLIP (VQAv2)	53.4	-
		BLIP	56.2	50.1
			B@4	**CIDEr**
Video Dialogue	AVSD	MTN (Le et al., 2019)	0.410	1.129
		PDC (Le et al., 2021)	0.429	1.194
		RLM (Li et al., 2021c)	0.459	1.308
		VGD-GPT	0.465	1.315

A public demo of LAVIS can be found at the temporary address: `http://34.123.225.190:8080/`