Appendix

The appendix includes

1. Details on dataset creation and statistics.
2. Implementation details for the various baselines.
3. Visualization of outputs.

A. Datasets

A.1. Creating Datasets for RAIV

We first discuss the creation of Im-Im, Im-Vid and Vid-Vid datasets which are aimed to have semantically rich representations.

RAIV tasks involve a pair of images/videos and a given statement to be classified as True or False. We create multiple datasets using existing vision-language datasets which contain SRL annotations, namely, ImSitu [47] and VidSitu [36]. The main reason for choosing datasets with SRL annotations is to obtain high-quality "action+object" information in the image or video. We first summarize these two datasets.

Briefly, the ImSitu dataset is created by first obtaining a set of verbs and their corresponding roles from FrameNet [33]. Then top image results are retrieved from the web which includes the particular verb, followed by a strict annotation pipeline to denote the various entities participating in the action. The VidSitu dataset, which serves as an extension of ImSitu to videos, obtains 10-second-long movie clips with multiple actions. Each video is then segmented into five 2-second events, with each segment annotated with a verb obtained from PropBank [16]. Then, a referring expression is used to denote the entities appearing in the videos, which are filled in the various roles.

For both ImSitu and VidSitu, we obtain the "object" information from an object detector. We utilize VinVL [49] which involves a FasterRCNN [31] trained on multiple object detection datasets OpenImages [18], COCO [22], Visual Genome [17] and Object365 [37], and then fine-tuned on Visual Genome.

We note that both ImSitu and VidSitu use different sets of verbs for annotations. Since our datasets include both images and videos, we simplify our setting by only utilizing verbs that are common to both datasets. While this reduces the total amount of available data, it hugely simplifies the dataset creation pipeline. We also prune verbs with less than 20 annotations in either dataset. This results in 243 verbs which are shared in both datasets.

Another issue arises in the semantic role labeling formats for the two datasets. ImSitu annotations are based on FrameNet [33] whereas VidSitu annotations are based on PropBank [16]. We use existing heuristics based on the ordering and the use of roles to map the SRLs from FrameNet to Propbank annotations. Since we are mostly concerned about the "action+object" setting and not the individual roles such as instruments or tools, noise in this conversion doesn't adversely affect the dataset quality. Further, the annotations for the entities in VidSitu have referring expressions or phrases describing the entity which is different from entity annotation in ImSitu containing only a single noun. We circumvent this issue by considering only the lemmatized noun for the referring expressions. We also avoid very common objects such as "person" which is usually associated with the agent performing an action.

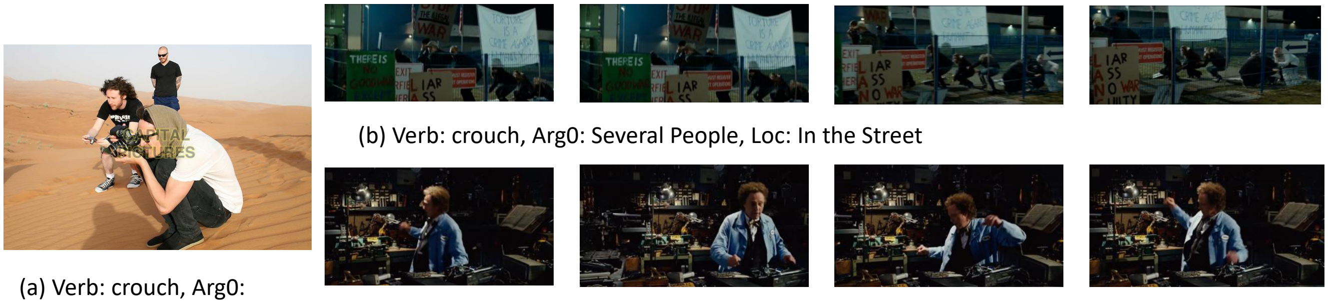
With both ImSitu and VidSitu datasets in hand, we now create RAIV datasets. We create the following variations: Image-Image (Im-Im), Image-Video (Im-Vid) and Video-Video (Vid-Vid) with images taken from ImSitu and videos taken from VidSitu. We note that while videos in VidSitu are 10 seconds long, for our experiments we only consider 2 second long clips which correspond to a particular event in the video. We further ensure the event is not duplicated in the next segment to avoid annotated entities not appearing within the given segment. After pruning, we are left with $63k$ images from ImSitu and $106k$ video segments from VidSitu. We utilize the same splits as in the original datasets to avoid any training dataset leakage into validation splits. For each of the datasets, we create approximately the same number of samples as in NLVR2 around $120k$ annotations with an even distribution of the verbs and objects but we note that our process allows creating more examples without any additional human effort.

We further take care to not introduce any spurious dataset bias. We follow NLVR2 in creating balanced validation and test sets by using the same unique statement where it is true for a particular pair and false for another pair in the given dataset to ensure no language-only bias in the dataset. The resulting datasets are suffixed with "T" to denote the statements are generated using templates resulting in Im-Im (T), Im-Vid (T), and Vid-Vid (T).

As our datasets are created semi-automatically, we also provide reasons for the false statements. For ease of evaluation we follow previous work in common-sense reasoning [19, 48] involving multiple-choice question setup where three reasons are provided and only one of the reasons is correct. The options are also generated via templates to prevent any language-only biases.

We summarize our pipeline for creating RAIV template datasets, i.e., Im-Im (T), Im-Vid (T), Vid-Vid (T) below.

1. Unify the annotations for ImSitu and VidSitu datasets, in particular the verbs.
2. Create mapping of objects, actions, and action+objects to image/video IDs in the datasets.
3. Sample a particular template based on object, action, or action + object. Then choose a particular object, action,



(a) Verb: crouch, Arg0: man, Loc: desert

(b) Verb: crouch, Arg0: Several People, Loc: In the Street

(c) Verb: dance, Arg0: A man in blue shirt, Loc: In the lab

Figure 4. Example creation of generating template-based queries.

action+object.

4. Choose a particular image/video satisfying the above criteria.
5. Choose two other image/video, one which satisfies and another which doesn't satisfy the criteria. This provides us with a True and False statement.
6. In previous step, choosing them at random makes the problem too simple, so we condition it on having at least one shared SRL such as verb, object or location.
7. For the false statement, provide the reason for being false.
8. Repeat the process until enough samples are obtained.

We illustrate this with an example in Fig 4. Suppose the chosen template was "action", "In both images, people are doing X" where "X" is the action. Assume the chosen action was "crouch". Let the first sample chosen be Fig 4 (a). Given this image, we choose a "true" video as in Fig 4(b) and "false" video (c). Further, for the "false" pair, we know both contain the verb crouch, so we can provide the reason "people crouch in I1 but not in I2.".

We note we restrict to limited possible templates yet covering a wide-variety of possibility based on whether it is "action", "object" or "action + object". The possible templates are:

1. "In both I1 and I2, {p1}."
2. "In at least one of I1 or I2, {p1}."
3. "In exactly one of I1 or I2, {p1}."
4. "In neither I1 nor I2, {p1}.

Here, {p1} is short for placeholder and {Image} refers to Image1 or Image2. We also note that the clause can be easily modified such as "In both I1 and I2, {p1}" is same as "{p1} in both I1 and I2". The placeholder {p1} depends on the type of template. For instance, if it is object, it is "Obj is present", for actions it is "Subj is performing Verb". These templates can then be used to get the reasoning in the form of: "In both I1 and I2", "In I1 but not in I2", "In I2 but not in I1" or "In neither I1 nor I2".

Note that for during training, the SRLs are obtained from a pre-trained SRL detection system on the provided captions such as [39].

For validation and test sets, we utilize all the available annotaitons. For instance, VidSitu provides 10 verb annotations for each segment. Thus, when comparing for same verb, we consider all 10 annotations. Similarly, for other SRLs. This makes our validation and test sets more robust to noisy ground-truth data.

A.2. Creating Natural Language Queries

As noted in main paper, we utilize LLMs to create Natural Language Queries. We note that there are both pros and cons of using natural language queries as opposed to template queries. The main advantage of templated queries is that the output sentence has very controlled information and as a result we can create a reasoning question directly from the template. However, such model is of little practical use.

On the other hand, natural language queries can be directly used by end-user but obtaining natural language queries via humans is prohibitively expensive. Instead, we opt to use natural language queries using LLMs. However, we note that use of LLMs can cause errors in the generated sentence and there is no easy way to rectify them. Further, the obtained LLM outputs cannot be used for reasoning.

To generate the queries, we use Vicuna-13B [50] model which is initialized from LLaMA [44] and trained on outputs from ChatGPT [26] a closed-source model by OpenAI.

We use the LLM in two ways: (i) to create Im-Im (G), Im-Vid (G) and Vid-Vid (G) which are generated counterparts to the original templated datasets introduced above (ii) to create IP2P dataset which is obtained from InstructPix2Pix. While used in similar ways, there are some key distinctions.

For Im-Im, Im-Vid and Vid-Vid datasets, we directly take all the visual input pairs, obtain their annotation information and pass it to the LLM and require it to generate a True statement. The obtained statement is then matched to another input pair for which it is false. Essentially, the “T” and “G” counterparts of the dataset have same visual input pairs but the exact sentences are different.

We prompt our LLM based on the original input query in the templated dataset. We use the following input:

""" Provide a True statement comparing the two images with the following information:

Image 1: {SRL} Image 2: {SRL}

The statement should be in the form of "{Template}, ...", only point out about {Image}. """

Here, {SRL} denotes the semantic roles for the given image/video, the {Template} denotes the chosen template as noted in previous section, and {Image} denotes which image was chosen (I1 or I2) for the true statement.

For instance, if the original query involved Fig4 (a), (c) with the template “In exactly one of”, with action+object, the input would be:

""" Provide a True statement comparing the two images with the following information:

Image 1: Verb: crouch, Subj: man, Loc: desert

Image 2: Verb: dance, Subj: a man in blue shirt, Loc: in the lab

The statement should be in the form of "In exactly one of the images, ...", only point out about image 1 """ This returns the output: "In exactly one of the images, a man is crouching in a desert." The same true statement is considered “false” for the other pair taken from corresponding “(T)” dataset.

For IP2P, the images are created using Stable-Diffusion. We have access to the image-caption pairs as well as the edit caption. To create a true statement, we provide the LLM with the original caption and the edit caption and ask it compare the images. To create a false statement, we keep the original caption but change the edit caption. Here, for each image pair we have a unique true and false statement.

To generate these queries:

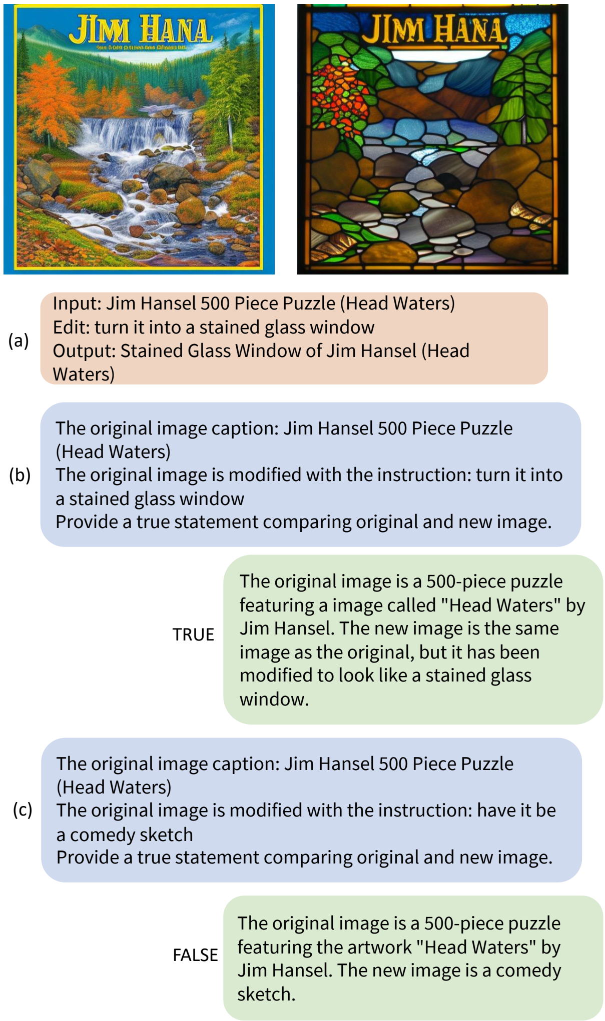
1. Choose a given image-pair from IP2P dataset. This has an associated original caption, edit caption, and an output caption.
2. Randomly sample another edit caption different from the given edit caption. Denote this as edit caption 2.
3. Use LLM to compare original caption, and edit-caption for True statement. Similarly, use original caption and edit caption 2 for False statement.

For generating output with a new caption we use the following prompt:

""" The original image caption: ...

The original image is modified with the instruction: ...

Provide a true statement comparing original and new image. """



Input: Jim Hansel 500 Piece Puzzle (Head Waters)
Edit: turn it into a stained glass window
Output: Stained Glass Window of Jim Hansel (Head Waters)

The original image caption: Jim Hansel 500 Piece Puzzle (Head Waters)
The original image is modified with the instruction: turn it into a stained glass window
Provide a true statement comparing original and new image.

The original image is a 500-piece puzzle featuring a image called "Head Waters" by Jim Hansel. The new image is the same image as the original, but it has been modified to look like a stained glass window.

The original image caption: Jim Hansel 500 Piece Puzzle (Head Waters)
The original image is modified with the instruction: have it be a comedy sketch
Provide a true statement comparing original and new image.

The original image is a 500-piece puzzle featuring the artwork "Head Waters" by Jim Hansel. The new image is a comedy sketch.

Figure 5. Illustration of generating queries for IP2P dataset. For a given image pair, we have (a) the sentence queries used to generate the image via Stable Diffusion (b) Using LLM to generate a True statement and (c) False statement by changing the edit caption.

As an example, to generate true statement for given pair : """ The original image caption: Jim Hansel 500 Piece Puzzle (Head Waters)

The original image is modified with the instruction: turn it into a stained glass window

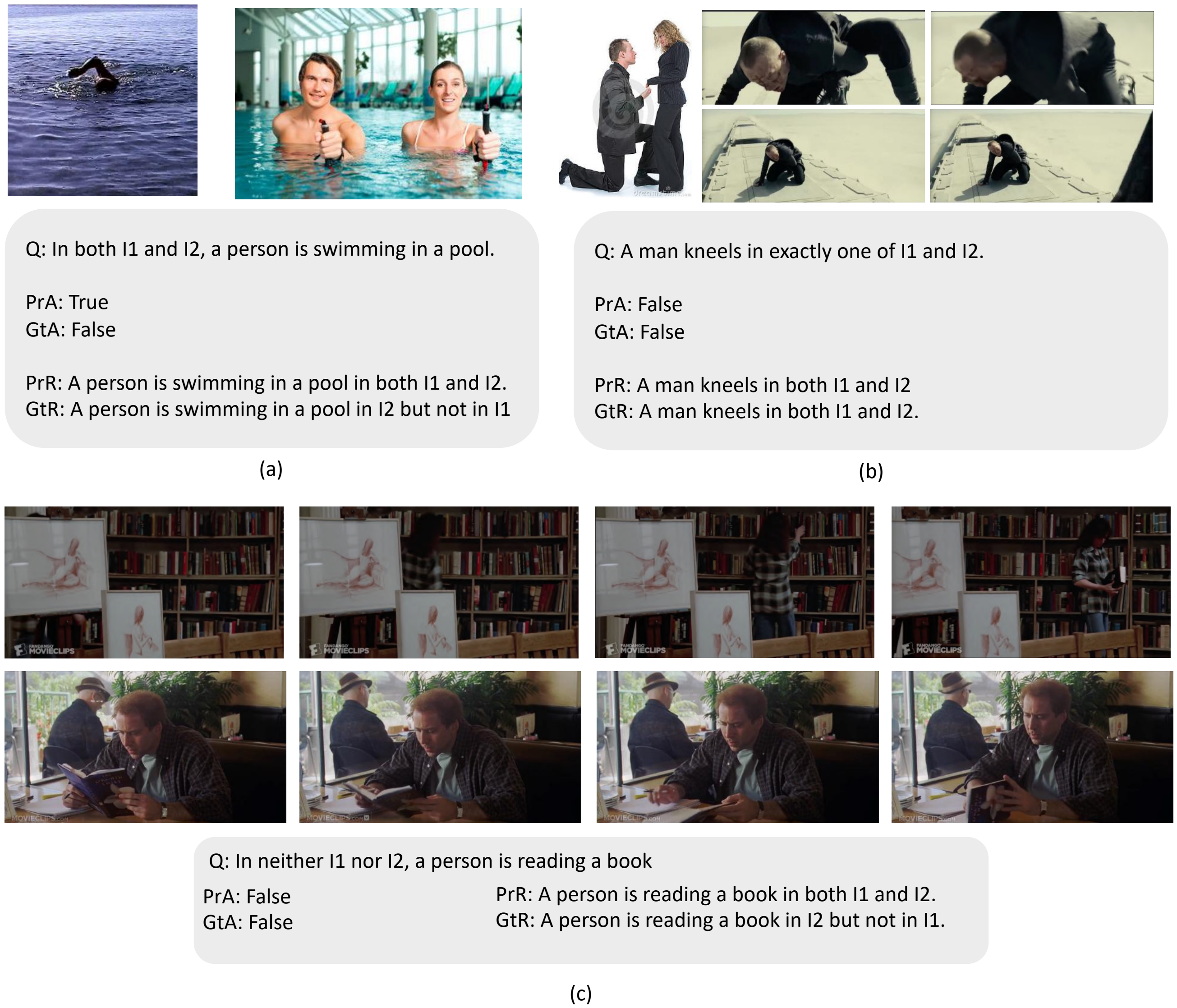
Provide a true statement comparing original and new image. """

Instead for a false pair:

""" The original image caption: Jim Hansel 500 Piece Puzzle (Head Waters)

The original image is modified with the instruction: have it be a comedy sketch

Provide a true statement comparing original and new image. """



Q: In both I1 and I2, a person is swimming in a pool.

PrA: True
GtA: False

PrR: A person is swimming in a pool in both I1 and I2.
GtR: A person is swimming in a pool in I2 but not in I1

(a)

Q: A man kneels in exactly one of I1 and I2.

PrA: False
GtA: False

PrR: A man kneels in both I1 and I2
GtR: A man kneels in both I1 and I2.

(b)

Q: In neither I1 nor I2, a person is reading a book

PrA: False
GtA: False

PrR: A person is reading a book in both I1 and I2.
GtR: A person is reading a book in I2 but not in I1.

(c)

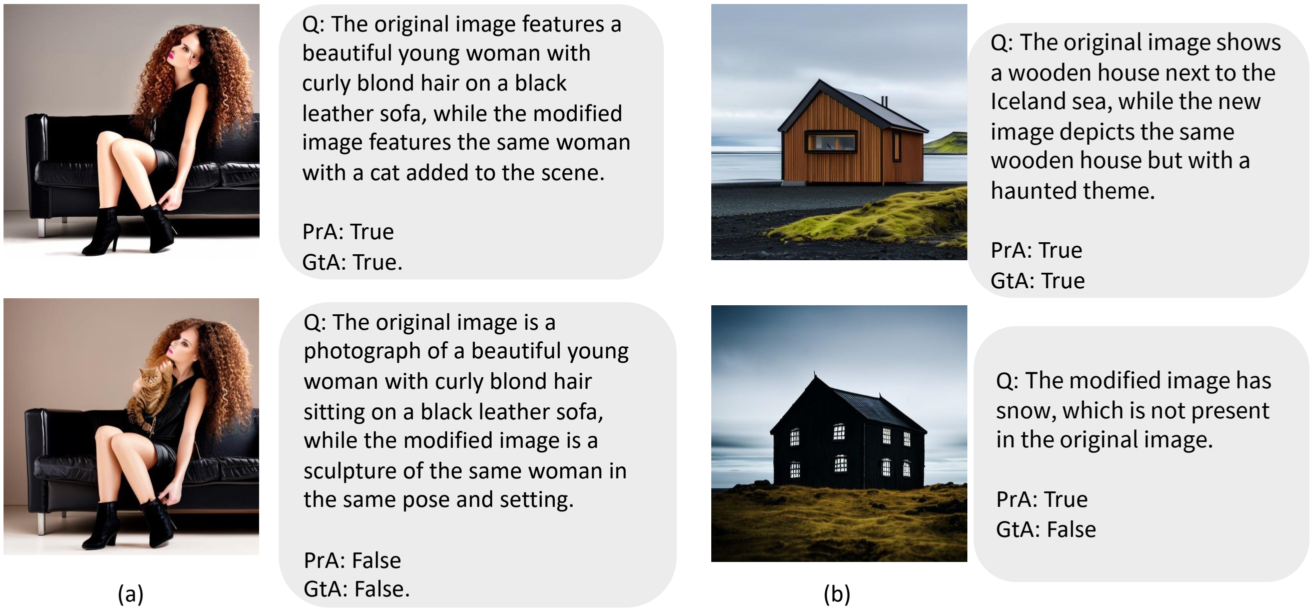
Figure 6. Model Predictions vs Ground-Truth for template-based ("T") validation datasets. (a) Im-Im (T), (b) Im-Vid (T), (c) Vid-Vid (T). PrA and GtA refer to Predicted and Ground-truth Answers respectively. PrR and GtR refer to predicted and ground-truth reasoning respectively.

B. Implementation Details

Implementation Details Our model and code are implemented in Pytorch. For all fine-tuning experiments, we follow identical settings as METER. For each dataset, we separately fine-tune the model for 10 epochs with differential learning rates of $1e^{-5}$ and $1e^{-4}$ for the bottom and top layers respectively.

We use 288×288 as the image dimension in all cases. For videos, we sample $K = 4$ frames per video where each video is 2 seconds long and sampled at 30 frames per second. For images, we simply provide a single temporal position embedding while for videos we have K temporal position embeddings. We use sinusoidal position embeddings following previous work [45].

In the task-specific pre-training step, we primarily use the COCO dataset instead of the entire ImgAll dataset in order to limit computation time, similar to the fine-tuning process on the downstream task. We also note that instead of using the object annotations available in COCO, we use the VinVL object detector outputs instead as it detects a larger number of categories outside of COCO. For videos, we use a subset of Kinetics videos from VATEX-en. We note that the videos in Kinetics are 10s long compared to 2s in the downstream dataset. To circumvent this issue, we first



Q: The original image features a beautiful young woman with curly blond hair on a black leather sofa, while the modified image features the same woman with a cat added to the scene.

PrA: True
GtA: True.

Q: The original image is a photograph of a beautiful young woman with curly blond hair sitting on a black leather sofa, while the modified image is a sculpture of the same woman in the same pose and setting.

PrA: False
GtA: False.

Q: The original image shows a wooden house next to the Iceland sea, while the new image depicts the same wooden house but with a haunted theme.

PrA: True
GtA: True

Q: The modified image has snow, which is not present in the original image.

PrA: True
GtA: False

Figure 7. Model Predictions vs Ground-truth for IP2P dataset. For a given pair of images, both the chosen True and False sentences are shown.

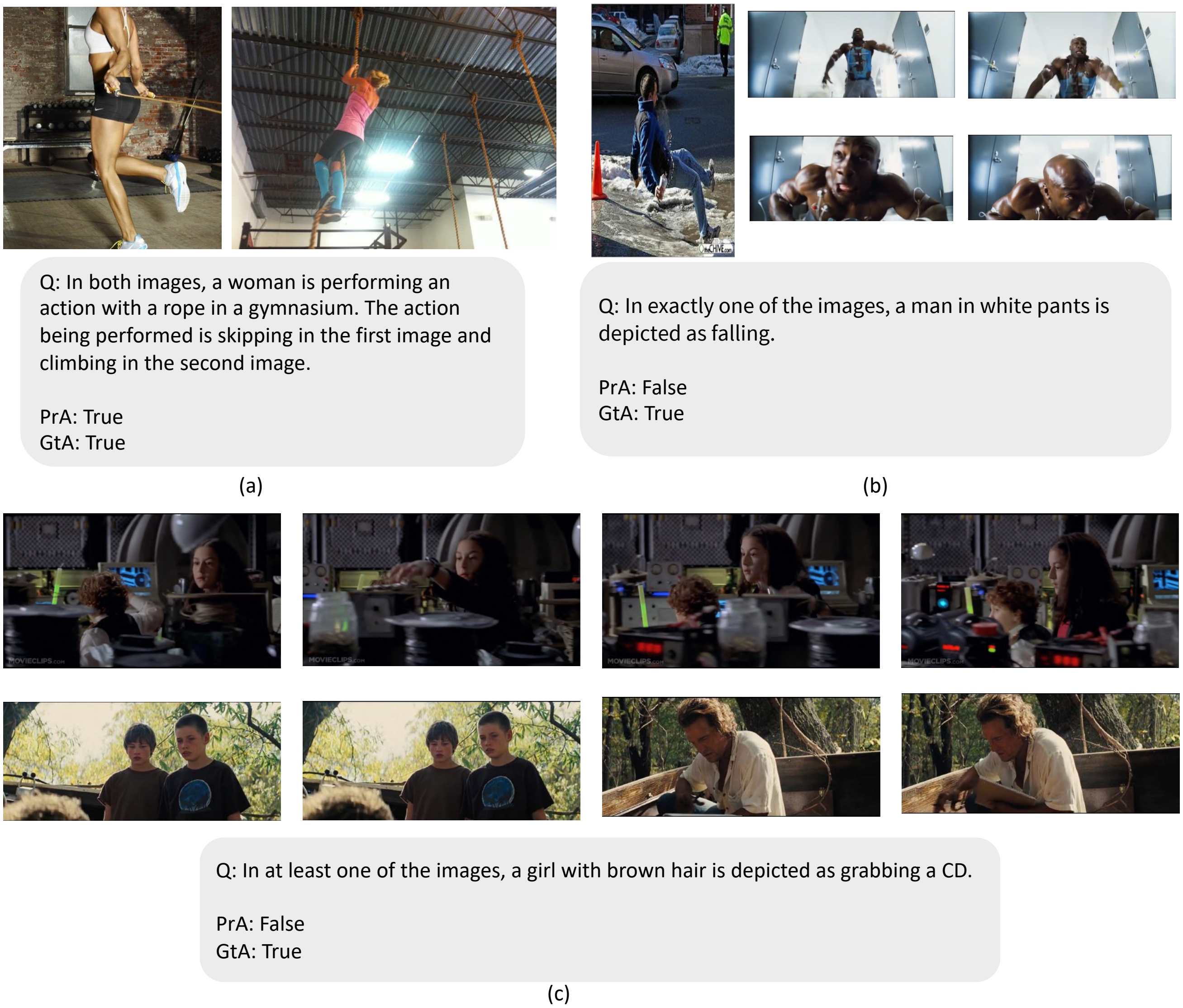
obtain an intersection of the videos from AVA-Kinetics [20] which gives us $5.7k$ videos where the keyframe of the person performing the action is provided. We particularly sample 2s clips around the keyframe. In general, we randomly sample 4 frames from the entire video.

We train for 10 epochs but reduce batch size to 256 with AdamW optimizer [24] with linear warm-up for initial 10% to $1e-4$ of the training followed by linear decay. We only utilize the last checkpoint and then perform fine-tuning on the target dataset. Most of our experiments are carried on 4x 2080Ti and 4x 3090Ti machines.

C. Visualization

We provide qualitative examples from our dataset and outputs of our model as follows:

1. On Template-based queries and Reasoning, namely, Im-Im (T), Im-Vid(T), Vid-Vid (T) in Figure 6
2. IP2P Generated queries in Figure 7
3. On Generated queries, Im-Im (G), Im-Vid(G), Vid-Vid(G) in Figure 8


Q: In both images, a woman is performing an action with a rope in a gymnasium. The action being performed is skipping in the first image and climbing in the second image.

PrA: True
GtA: True

(a)

Q: In exactly one of the images, a man in white pants is depicted as falling.

PrA: False
GtA: True

(b)

Q: In at least one of the images, a girl with brown hair is depicted as grabbing a CD.

PrA: False
GtA: True

(c)

Figure 8. Model Predictions vs Ground-Truth for generated queries ("G") validation datasets. (a) Im-Im (G), (b) Im-Vid (G), (c) Vid-Vid (G). PrA and GtA refer to Predicted and Ground-truth Answers respectively.

References

[1] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra. Vqa: Visual question answering. *International Journal of Computer Vision*, 123:4–31, 2015. 1

[2] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1708–1718, 2021. 3

[3] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*, 2022. 2, 5

[4] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 381–389, 2018. 3

[5] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *ArXiv*, abs/1504.00325, 2015. 1

[6] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *ECCV*, 2020. 1, 2, 3

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

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. 2

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 2, 7

[11] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Nanyun Peng, Zicheng Liu, and Michael Zeng. An empirical study of training end-to-end vision-and-language transformers. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18145–18155, 2022. 1, 2, 3, 5, 6

[12] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. *International Journal of Computer Vision*, 127:398–414, 2017. 1

[13] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021. 2

[14] Zeeshan Khan, C.V. Jawahar, and Makarand Tapaswi. Grounded video situation recognition. *ArXiv*, abs/2210.10828, 2022. 3

[15] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*, 2021. 1, 2

[16] Paul R Kingsbury and Martha Palmer. From treebank to propbank. In *LREC*, pages 1989–1993, 2002. 9

[17] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. 2016. 6, 9

[18] Alina Kuznetsova, Hassan Rom, Neil Gordon Alldrin, Jasper R. R. Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4. *International Journal of Computer Vision*, 128:1956–1981, 2020. 9

[19] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L. Berg. Tvqa: Localized, compositional video question answering. In *EMNLP*, 2018. 6, 9

[20] Ang Li, Meghana Thotakuri, David A. Ross, João Carreira, Alexander Vostrikov, and Andrew Zisserman. The ava-kinetics localized human actions video dataset. *ArXiv*, abs/2005.00214, 2020. 6, 13

[21] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven C. H. Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, 2021. 1, 2, 3, 6, 7

[22] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 5, 6, 9

[23] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 6

[24] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 6, 13

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

[26] OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023. 2, 5, 10

[27] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24, 2011. 6

[28] Sarah Pratt, Mark Yatskar, Luca Weihs, Ali Farhadi, and Aniruddha Kembhavi. Grounded situation recognition. In *ECCV*, 2020. 3

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2, 6

[30] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. 2

[31] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1137–1149, 2015. 7, 9

[32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 2, 5

[33] Josef Ruppenhofer, Michael Ellsworth, Myriam Schwarzer-Petruck, Christopher R Johnson, and Jan Scheffczyk. Framenet ii: Extended theory and practice. Technical report, International Computer Science Institute, 2016. 9

[34] Arka Sadhu, Kan Chen, and Ram Nevatia. Video object grounding using semantic roles in language description. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[35] Arka Sadhu, Kan Chen, and R. Nevatia. Video question answering with phrases via semantic roles. In *NAACL*, 2021. 3

[36] Arka Sadhu, Tanmay Gupta, Mark Yatskar, Ram Nevatia, and Aniruddha Kembhavi. Visual semantic role labeling for video understanding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021. 1, 3, 4, 5, 9

[37] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8430–8439, 2019. 9

[38] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia, July 2018. Association for Computational Linguistics. 6

[39] Peng Shi and Jimmy Lin. Simple bert models for relation extraction and semantic role labeling. *arXiv preprint arXiv:1904.05255*, 2019. 3, 4, 5, 6, 10

[40] Emma Strubell, Patrick Verga, Daniel Andor, David Weiss, and Andrew McCallum. Linguistically-informed self-attention for semantic role labeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5027–5038, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics. 4

[41] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*, 2020. 2

[42] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428, Florence, Italy, July 2019. Association for Computational Linguistics. 1, 3

[43] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019. 2, 3

[44] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2, 5, 10

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 6, 12

[46] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 6

[47] Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. Situation recognition: Visual semantic role labeling for image understanding. In *Conference on Computer Vision and Pattern Recognition*, 2016. 1, 3, 4, 5, 9

[48] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6713–6724, 2019. 6, 9

[49] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Making visual representations matter in vision-language models. *CVPR 2021*, 2021. 3, 7, 9

[50] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. 5, 10