Supplementary Material
Movies2Scenes: Using Movie Metadata to Learn Scene Representation

Shixing Chen Chun-Hao Liu Xiang Hao Xiaohan Nie Maxim Arap Raffay Hamid
Amazon Prime Video
{shixic, chunhaol, xianghao, nxiaohan, maxarap, raffay}@amazon.com

In this supplementary material, we provide two sections to better support the arguments and results in the main paper: (a) more details of the specific settings in our experiments for better reproducibility, and (b) more extensive qualitative results to better demonstrate the interpretability of our learned representations. The sections will be presented with information corresponding to different datasets used in the main paper including: Movie Contrastive Learning 30K (MovieCL30K) dataset, Long-Form Video Understanding (LVU) dataset [18], MovieNet dataset [12] [17] and Mature Content Dataset (MCD).

1. Experiment Details

We use PyTorch 1.8 [15] as our deep learning library and NVIDIA Tesla A100/V100 GPUs for computation. During contrastive learning, we use 8 GPUs with distributed data and model parallelism. For supervised learning of MLP on downstream tasks, only 1 GPU is needed.

1.1. MovieCL30K

This section corresponds to §4.1.2 in the main paper.

a. Shot-encoder: Our shot encoder has two key differences comparing with ShotCoL [2]. First, we can select the positive keys during training, which makes training more efficient where we do not use the stale positive keys for epochs before updating them. Second, ShotCoL [2] focused on learning a representation that is most useful for the scene boundary detection task, so it could benefit from contextual and semantic information in a neighborhood size similar to the length of a scene. However, we observed that when the neighborhood size is relatively large (e.g., 16) as selected in ShotCoL, the positive key may end up being almost identical to the query. This is still useful information for scene boundary detection task because there could be almost identical shots in a scene. However, for our objective to learn a representation that focuses on appearance, this may reduce the effectiveness of representations because the positive key is more similar to augmented images which were demonstrated to be less effective in [2].

b. Movie-level similarity learning: When using movie metadata to train the movie encoder (Figure 2 in the main paper) on MovieCL30K, we used SGD to optimize with a learning rate of 0.1, batch size of 256 and epoch number of 100. The same set of hyper-parameters was applied to all three types of movie metadata (co-watch, genre, and synopsis). Recall that within each batch, there are 256 pairs of movies represented by feature matrices extracted from our shot encoder, and the dimension of each feature matrix is 1024×512. These pairs are passed through $\mathbf{E}_{\text{movie}}$ to predict whether two movies are similar based on movie metadata.

c. Scene contrastive learning: After movie-level similarity learning, we select the set of similar scene-pairs based on the learned space in $\mathbf{E}_{\text{movie}}$. Specifically, after extracting features of all shots in each input movie by $\mathbf{E}_{\text{shot}}$, the movie is represented as a matrix of $\text{M}\times512$, where M is the number of shots in the movie. Notice that the length of the movie is no longer restricted to 1024, so that all shots can be considered during similar scene selection. The feature matrix is then passed to $\mathbf{E}_{\text{movie}}$ before the last fully-connected layer, and becomes a new feature matrix. Similar process is done on another movie considered similar to the input movie with N shots, and the shot adjacency matrix $\mathbf{A}$ of these two movies takes the size of $\text{M}\times\text{N}$. We then go through all 9×9 windows in $\mathbf{A}$ with stride 1, and calculate the average value in each window to represent the scene-level similarity of the two movies. Finally, we pop the scene-pairs with top 50% highest scene-level similarity scores while keeping the selected scenes to be non-overlapping. This generates a set of scene-pairs for each type of movie metadata.

With the generated set of scene pairs, we use them for scene contrastive learning following the MoCo framework [9], while substituting encoder to be ViT [4] and optimization to be AdamW [14] instead of SGD [8]. Specifically, following [9], we use feature dimension of 128, queue size of $65,536$, MoCo momentum of 0.999 and softmax temperature of 0.07 during momentum contrastive learning. Following [4] [3], we use learning rate of $1.5\text{e}{-4}$, weight decay of 0.01, number of warm-up epochs of 40, batch size of 128 and number of epoch of 100.

Hyper-parameters	Classification							Regression	
	place	director	relation	speak	writer	year	genre	view	like
learning rate	0.1	1.0	0.01	0.1	0.5	0.1	0.1	0.01	0.1
batch size	128	256	16	256	32	128	256	16	16
epoch	400	400	400	400	400	400	400	500	500
dropout	0.1	0.5	0.5	0.1	0.25	0.75	0.1	0.8	0.5

Table 1. Hyper-parameters used in the MLPs for LVU tasks.

1.2. LVU

When producing the results in Table 2 in the main paper, following [18] where the model of each task is trained separately with parameters selected by validation set, we selected the parameters and hyper-parameters of MLP for each task by the validation set in LVU and presented corresponding results on test set. The hyper-parameters used in MLPs of each task on representations pre-trained by co-watch is shown in Table 1.

1.3. MovieNet

This section corresponds to §4.3 in the main paper.

a. Place tagging: For the results of Ours in Table 3 in the main paper, we used MLP with two 512-dimensional hidden layers optimized by SGD with leaning rate of 5.0, dropout of 0.25, epoch number of 200 and batch size of 512. The problem was formulated as a multi-label classification task and optimized by BCEWithLogitsLoss [15].

b. Scene boundary detection: For the results of Ours in Table 4 in the main paper, we used MLP with two 512-dimensional hidden layers optimized by SGD with leaning rate of 0.03, dropout of 0.8, epoch number of 800 and batch size of 4096.

1.4. MCD

We used SlowFast 8x8 R50 and SlowFast 8x8 R101 for the SlowFast models [6] used in Table 5 in the main paper. Both models take 64 frames from each video clips. When extracting representation from the SlowFast models, we used the average pooling layers before the final classification layer, and the representation has 2304 dimensions concatenated from the slow and fast pathways. Similarly for the X3D-L model [5], we used the fully connected layer before the final classification layer, which has 2048 dimensions, and the X3D-L model takes 16 frames from each video clips. For the CLIP model [16], we used ViT-B/16 based visual encoder, and it takes the same 9 frames as the input to our model from each video clip. We first extracted the embeddings with 512 dimensions from each frame and then do an average pooling across all 9 frames, which is used as the representation for the CLIP model. For training the model to classify the age-appropriate activities, we used a 3-layer MLP model with 512 nodes in the hidden layers.

2. Additional Results

2.1. MovieCL30K

We present the similar scene pairs selected by our scene representation learned on co-watch in Figure 1. We also present the similar scene-pairs found by pre-trained CLIP visual features [16] in Figure 2 and the ones found by pre-trained Merlot Reserve visual features [19] in Figure 3, respectively. We can see that scene-pairs found by our approach are significantly more **thematically similar**, while the ones found by other features focus much more on **appearance similarity**. Moreover, CLIP [16] and Merlot Reserve [19] features produce results that are mostly related to human faces, which is not sufficiently useful for general-purpose semantic scene understanding. These observations provide insights about the effectiveness of our approach on a wide variety of downstream tasks related to semantic scene understanding compared to other state-of-the-art representations. Lastly, we can also noticed that the scenes pairs found by Merlot Reserve visual features [19] are similar to the ones found by CLIP visual features [16], which can indicate that the added audio modality in Merlot Reserve in not directly influencing the distribution of embedding in visual encoder before the modalities are fused.

2.2. LVU

To demonstrate the effectiveness of our learned scene-representation, we use the place-labeled scenes from LVU data [18] in a retrieval-setting. Specifically, using query scenes each with a particular place-label from the validation-set, we retrieve 1, 5, and 10 nearest neighbors from the training-set using their L_2 distances. Precision results for various settings are given in Table 2 where our encoder is compared with the pre-trained visual encoder of CLIP [16] and the SlowFast model [6] pre-trained on Kinetics [1] and AVA [13].

Moreover, to provide qualitative insights into the effectiveness of our learned scene-representation, Figure 4 shows the retrieval results using an example query for 4 of the 6 categories based on ours as well as CLIP [16] visual representation. It can be seen that although CLIP visual representation can capture local appearance-based patterns effectively, it is not able to capture longer-duration semantic aspects of scenes. In contrast, our representation is able to

Figure 1. **Similar scene-pairs found by our representation -** Given a similar movie-pair *Need For Speed* and *The Fast & Furious: Tokyo Drift*, we present representative examples from the set of similar scene-pairs (connected by orange arrow) found by our representation (sorted by similarity). Comparing with the ones found by CLIP visual feature [16] in Figure 2 and Merlot Reserve feature [19] in Figure 3, these scene-pairs are more thematically meaningful, which contributes to the effectiveness of our representation on downstream tasks related to semantic scene understanding.

Figure 2. **Similar scene-pairs found by CLIP -** Given a similar movie-pair *Need For Speed* and *The Fast & Furious: Tokyo Drift*, we present representative examples from the set of similar scene-pairs (connected by orange arrow) found by CLIP visual feature [16] (sorted by similarity). Comparing with the ones found by our representation in Figure 1, these scene-pairs focus more on appearance-based similarity and are mostly related to human faces, which are not sufficiently semantically meaningful for semantic scene understanding.

Figure 3. **Similar scene-pairs found by Merlot Reserve -** Given a similar movie-pair *Need For Speed* and *The Fast & Furious: Tokyo Drift*, we present representative examples from the set of similar scene-pairs (connected by orange arrow) found by Merlot Reserve visual feature [19] (sorted by similarity). Comparing with the ones found by our representation in Figure 1, these scene pairs are more similar to the ones found by CLIP in Figure 2, which focus more on appearance-based similarity and are mostly related to human faces, which are not sufficiently semantically meaningful for semantic scene understanding.

Models	CLIP [16]			SlowFast [6]			Ours		
Architecture	ViT-B/16 [4]			ResNet-101 [11]			ViT-B/16 [4]		
Pre-training data	400M image-text pairs			Kinetics [1]+AVA [13]			MovieCL30K		
Pre-training task	image-text similarity			action recognition			scene contrast		
	top-1	top-5	top-10	top-1	top-5	top-10	top-1	top-5	top-10
office	30	20	19	0	16	17	30	36	26
airport	0	15	10	0	0	10	0	15	10
school	28.57	26.66	25	76.19	20.47	32.14	50	49.04	45.95
hotel	35.71	28.57	29.28	0	0	10	42.85	20	24.28
prison	52.94	42.94	42.94	35.29	40	20.29	58.82	45.88	40
restaurant	20	13.99	13.99	0	20	10	40	16	14
all queries	35.08	29.64	28.85	38.59	22.63	21.84	**47.36**	**39.29**	**35.70**

Table 2. The place-labeled scenes in LVU data [18] are formulated to a retrieval setting, where the goal is to retrieve similar scenes from training-set given query scenes from validation-set. Representations pre-trained on different configurations as specified in the table are compared. The precision results for different place categories are reported with the size of retrieved set to be 1, 5 and 10.

Figure 4. Qualitative results of place retrieval using LVU data [18] are shown. For each query scene in validation-set, two similar scenes from training-set are retrieved based on ours and CLIP visual representation [16]. Example results show that our feature can capture both scene-appearance as well as their broader thematic signature, while CLIP [16] can only capture scene-appearance effectively.

capture the appearance as well as semantic aspects of scenes effectively, and is therefore able to avoid the types of confusions that confound the CLIP representation.

We applied average pooling instead of concatenation on CLIP feature [16] because during retrieval, we do not want the order of shots to influence the results, thus, having one vector per input scene is reasonable. Notice that for the place of airport, both our representation and CLIP have top-1 accuracy of 0, it is because the number of query is very limited in the validation set of LVU [18] (4 airport-labeled scenes), and thus the airport example presented in Figure 4 comes from the top-2 retrieval result.

In Figure 5, we show five example scenes with their ground truth genre in LVU as well as our prediction. We can see that our model was able to capture the feature that is useful for correct genre prediction in most cases. For cases like example 5 on the last row, we assume it is because it is sometimes difficult to identify genre from just one scene in the movie, and since the meta information like genre in LVU was retrieved from IMDB entries [18], they can not always reflect the genre of a specific scene.

In Figure 6, we show six examples from the way-of-speaking prediction task in LVU. We can see that when the visual information is sufficient to make predictions, our model can perform well, but for cases like example 6 on the last row, it can be insufficient to predict just based on visual cues, and this might indicate that for tasks like way-of-speaking prediction, it may be beneficial to include audio modality for better accuracy. This also corresponds to the results in Table 2 of the main paper, where the accuracy of way-of-speaking is lower than other tasks. For relationship prediction in Figure 7 we can see that the model can make ambiguous predictions when there are more than one pairs of characters in the scene, and this may indicate that when analyzing relationship in scenes, it can be helpful to focus more on leading characters.

Figure 5. Additional qualitative results on LVU datasets [18]. Five example scenes with their ground truth genre labels as well as our predictions are shown. For example 5, our prediction is different from the label even though looking only at the visual content of this particular scene it makes sense to infer it as a romantic scene. Our hypothesis is that as the genre label of the LVU dataset was acquired from movie-level meta data, sometimes it is not directly applicable to the genre of all of the constituent scenes of a movie.

2.3. MovieNet

We present qualitative results on MovieNet [12] [17] dataset in Figure 8. This corresponds to Table 3 and Table 4 in the main paper and includes examples on place tagging as well as scene boundary detection (SBD) tasks. We show three examples from test set of MovieNet, and in each example, there are two scenes divided by the green dotted line. For each scene, the task is to predict what are the multi-label place tags of the scene, and for the two scenes together, the goal is to predict whether the shot boundaries between each pair of shots are also scene boundaries.

We can see that for SBD, our model can perform well to clearly identify the scene boundaries. For place tagging, it is a much more difficult task involving holistic understanding of the scenes, and although our model outperformed existing state-of-the-art models by a large margin in Table 3 in the main paper, it is still a really challenging and unsolved

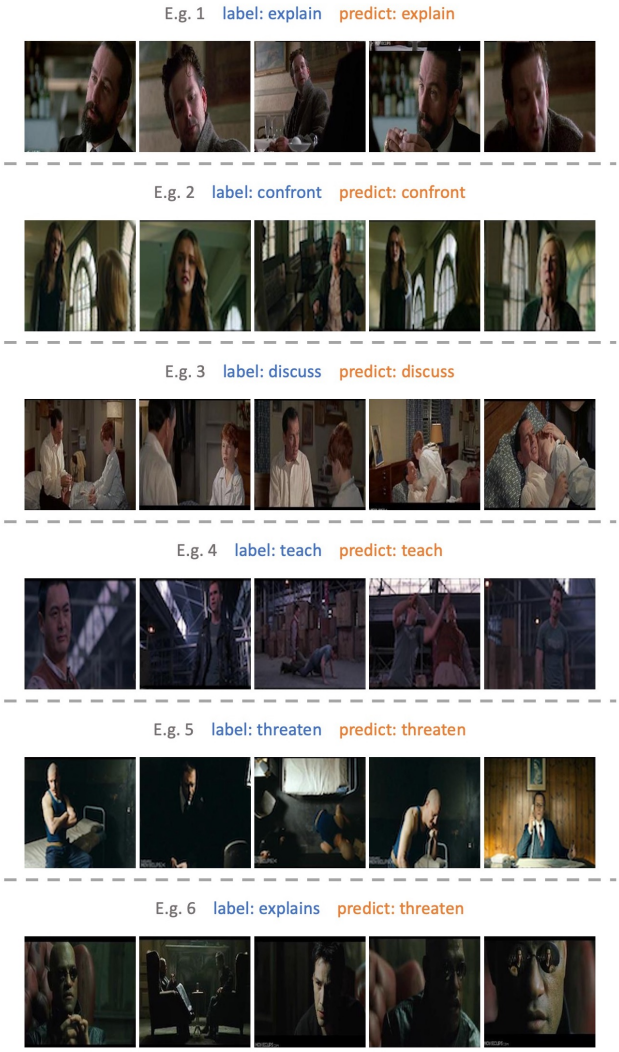

Figure 6. Six examples from the way-of-speaking prediction task in LVU [18]. Results show that when visual information is sufficient to predict way-of-speaking, our method can perform well, and for cases like example 6, it might be beneficial to add audio modality to further improve the accuracy.

task. For example, some places are not easy to identify based on a few frames (e.g. playground), and some places can vary a lot in term of appearance but have same place tag (e.g. car). This is also partially caused by the lack of labeled data, where for the 90-category multi-label problem, there are 19.6K place tags, with ∼11.7K for training, leading to ∼130 labeled training tags per category on average.

2.4. MCD

Representative examples from the three age-appropriate activities in MCD dataset are provided in Figure 9. We also show the samples for each of the 4 classes of our MCD dataset in Figure 10 along with the corresponding detection results (i.e., the class with maximum probability) from both our model and CLIP model. For sex examples, we

Figure 7. Examples of relationship prediction task in LVU [18]. Sometimes the relationship can be ambiguous when there are more than two leading characters in the scene. Also, the uncertainty of some relationships makes prediction even more challenging. For example, it can be difficult to distinguish husband wife and boyfriend girlfriend without semantically understanding of context and plot.

Figure 8. Qualitative results on MovieNet place tagging [12] and scene boundary detection tasks [17]. Although our model outperformed existing state-of-the-art models by a large margin in Table 3 in the main paper, multi-labeled place tagging is still a really challenging problem. Some tags may not be apparent and the intra-category variance is large in this task.

Figure 9. Examples of 3 types of age-appropriate activities in our data. Sensitive parts of images have been intentionally redacted here. See supplementry materials for more examples.

Figure 10. Representative examples in MCD comparing the predictions based on our representation with CLIP visual feature. Sensitive parts of some images are intentionally redacted here.

Models	Pre-training data	sex	violence	drug-use	average
ShotCoL [2]	movie shot pairs	62.3	58.7	47.1	56.0
MerlotReserve [19]	Youtube videos	77.0	68.2	53.4	66.2
BridgeFormer [7]	image+video	74.8	61.4	61.2	65.8
Ours	MovieCL30K	**81.5**	**70.2**	**61.8**	**71.1**

Table 3. Comparisons on MCD with other pre-trained models.

can see that our model has higher confidence scores compared to CLIP model especially when the images have dark illumination. For violence, the CLIP model sometimes mistakes scenes with two closed persons as sex as shown in the first example. Similarly for drug-use examples, our model classifies them more confidently, and CLIP model misses the small cigarette in the third example and classifies it as none. The none examples show that the CLIP model often mistakes them with other classes, such as violence and drug-use, while our model is able to classify them correctly. This indicates our scene representation performs better than CLIP on age-appropriate activities, demonstrating the effectiveness of our representation in video moderation. Additional quantitative results on MCD are provided in Table 3 for comparisons with other pre-trained large models.

3. Additional Insights

a. Details of scene adjacency matrix: Consider an example movie-pair $\mathbf{x_1}$ and $\mathbf{x_2}$ with $\mathbf{m}$ and $\mathbf{n}$ number of scenes respectively. The shape of their scene adjacency matrix $\mathbf{B}$ is $\mathbf{m}\times\mathbf{n}$. Each value in $\mathbf{B}$ indicates the similarity score between two scenes, one from each movie. We rank all the similarity scores in $\mathbf{B}$ so that we know which pairs of scenes are most similar. We first select the scene-pair (say $\mathbf{m}_0$ and $\mathbf{n}_0$) in $\mathbf{B}$ with the highest similarity score and add it to $\mathbf{P}_{\text{scene}}$. Moreover, we keep a record of this scene-pair, so that $\mathbf{m}_0$ and $\mathbf{n}_0$ will not be selected again from movie-pair $\mathbf{x_1}$ and $\mathbf{x_2}$. We then move on to the scene-pair corresponding to the second highest value in $\mathbf{B}$, and only add it to $\mathbf{P}_{\text{scene}}$ if neither of the scenes in that scene-pair is $\mathbf{m}_0$ or $\mathbf{n}_0$. We carry out this process for the top 50% of the most similar scene-pairs in $\mathbf{B}$. This movie-pair level routine is repeated for all pairs of movies in our dataset to build $\mathbf{P}_{\text{scene}}$.

b. Effectiveness of pre-trained CLIP weights: In general, it has been demonstrated that pre-training does not always result in improvements [10]. Specifically for our case, there are two key reasons why using pre-trained CLIP does not offer additional benefits. First, the domain gap between pre-trained CLIP (internet images and texts) and our data (movies) is quite high. Second, CLIP uses individual images and incorporates no temporal information. However, our use of scenes heavily relies on information among frames and shots.

References

[1] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arXiv:1808.01340*, 2018. 2, 6

[2] Shixing Chen, Xiaohan Nie, David Fan, Dongqing Zhang, Vimal Bhat, and Raffay Hamid. Shot contrastive self-supervised learning for scene boundary detection. In *CVPR*, 2021. 1, 8

[3] Xinlei Chen*, Saining Xie*, and Kaiming He. An empirical study of training self-supervised vision transformers. *arXiv preprint arXiv:2104.02057*, 2021. 1

[4] 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. 1, 6

[5] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *CVPR*, 2020. 2

[6] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, 2019. 2, 6

[7] Yuying Ge, Yixiao Ge, Xihui Liu, Dian Li, Ying Shan, Xiaohu Qie, and Ping Luo. Bridgeformer: Bridging video-text retrieval with multiple choice questions. In *CVPR*, 2022. 8

[8] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv:1706.02677*, 2017. 1

[9] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 1

[10] Kaiming He, Ross Girshick, and Piotr Dollar. Rethinking imagenet pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 9

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[12] Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. Movienet: A holistic dataset for movie understanding. In *ECCV*, 2020. 1, 7, 8

[13] Ang Li, Meghana Thotakuri, David A Ross, João Carreira, Alexander Vostrikov, and Andrew Zisserman. The ava-kinetics localized human actions video dataset. *arXiv:2005.00214*, 2020. 2, 6

[14] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 1

[15] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 1, 2

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv:2103.00020*, 2021. 2, 3, 4, 6

[17] Anyi Rao, Linning Xu, Yu Xiong, Guodong Xu, Qingqiu Huang, Bolei Zhou, and Dahua Lin. A local-to-global approach to multi-modal movie scene segmentation. In *CVPR*, 2020. 1, 7, 8

[18] Chao-Yuan Wu and Philipp Krahenbuhl. Towards long-form video understanding. In *CVPR*, 2021. 1, 2, 6, 7, 8

[19] Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. Merlot reserve: Multimodal neural script knowledge through vision and language and sound. In *CVPR*, 2022. 2, 3, 5, 8