

Supplementary Materials

ANONYMOUS AUTHORS

1 RELATED WORK SUPPLEMENT

Casual Inference. Some researchers focus on human-centered causal inference [7, 8, 18]. [8] aims to detect causal utterances for a non-neutral targeted utterance in conversations. It proposes a novel usage of conversation graphs, incorporating emotional information and social commonsense knowledge, to improve the detection of causal utterances with different emotions. [18] proposes the Knowledge-Bridged Causal Interaction Network (KBCIN) for identifying causal utterances responsible for a non-neutral emotion in conversations. It leverages commonsense knowledge as three bridges: semantics-level bridge, emotion-level bridge, and action-level bridge. This work constructs conversational graphs, and utilizes CSK to capture inter-utterance dependencies and reason the target emotion. [11] proposes a retroductive reasoning task different from our task. It requires the model to generate the caption to explain the observation reasoning. Compared with [11], our task has two characteristics: (1) Our task aims to retrieve the target video to explain the textual observation. (2) Our task requires the AI system to complete the middle action flow between the video and observation. Towards different task requirements, the annotations of our dataset are designed with specialization.

Text-video Retrieval. Some researchers apply the reasoning method for solving the text-video retrieval task [1–3]. [1] presents the Hierarchical Graph Reasoning (HGR) model for fine-grained video-text retrieval. It decomposes the matching process into global-to-local levels using a hierarchical semantic graph, generating hierarchical textual embeddings for improved cross-modal matching. [2] proposes a two-stream framework for video-text retrieval, incorporating concept information and semantic-level matching. It introduces concept propagation using a commonsense graph to enrich video representations, achieving improved performance on benchmark datasets.

Cross-modal Transformer. Some researchers introduce transformer-based architectures into the cross-modal reasoning tasks [5, 13, 19]. [13] presents an empirical investigation of the transformer guided chaining approach for multi-hop First-Order Logic reasoning. The developed reference implementation, Chainformer, analyzes accuracy, generalization, interpretability, and performance, providing insights into the strengths, limitations, and future research directions in logic reasoning. [19] introduces the Multimodal Commonsense Transformer (MCOMET), a unified framework for physical audiovisual commonsense reasoning. MCOMET leverages higher-ordered temporal relationships across modalities, and employs a feature collection and propagation mechanism for selective cross-modal flow, which demonstrates superior performance on the PACS benchmark.

2 DATASET SUPPLEMENT

2.1 Annotation Pipeline

We visualize the annotation pipeline of the Tex-COIN dataset in the figure 1.

2.2 Data Collection

Before illustrating the annotation process, we introduce how to collect the data in our Tex-COIN. Specifically, The data in our Tex-COIN dataset comes from two sources: (A) We manually crop 10,000 video clips from 92 well-known TV shows, including Mr. Bean, Grey’s Anatomy, etc. In addition, we carefully select 10,000 video clips from the previous

Author’s address: Anonymous Authors.

Action	Class
Intransitive verb	(dance, applaud, stand up, run, sneeze, cook, work, bend over, write, sit down, walk, lie down, answer phone, smoke, squat down, fall down, crawl, awake, get up, turn around, clean, yawn, jump, speak)
Transitive verb	(touch, kiss, hit, ride, grab, put down, hug, fix, throw, eat, drink, pour, wash, pick up, hold, open, close, make up, give something to somebody, take something from somewhere, move, take on, take off, put something into/on somewhere, cover, walk into, walk out, wipe, flip, tidy, watch, search, tie, feed, untie, play)
Object	(bag, bed, blanket, book, box, cleaning supplies, cabinet, chair, clothes, cup, dish, door, door-knob, floor, food, laptop, light, mirror, paper, phone, picture, refrigerator, shelf, shoe, sofa, table, television, towel, window, body, hair, drink, pillow, tableware, lid, bottle, bowl, face, hat, coat hanger, plug, stationery, bucket, pots, pool, switch, toy, washing machine, toiletries, kitchen utensils, rope, repair tool, key, curtain, camera, remote control, mouse, battery, water dispenser, lighter, wall, railing, glasses, rubbish, earphone, pet, musical instrument, fitness equipment, watch, dryer, toilet, cigarette, seasoning, unknown, tube, umbrella, plant, transportation, baby)

Table 1. Statistics of action commonsense knowledge for the Tex-COIN dataset.

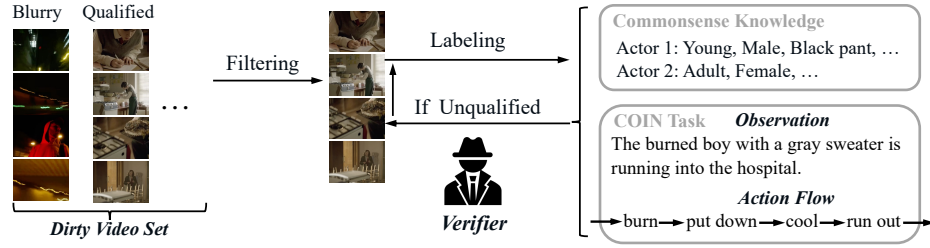


Fig. 1. A diagram of the dataset construction pipeline.

datasets, including the HC-STVG dataset [16], the ava-actions dataset [4], and the TO-MAR dataset [10]. (B) 20,000 lifestyle videos are carefully collected from the Charades dataset [15] and the vidor dataset [14]. Not all the collected videos are suitable for annotation in the next step, like the videos with extremely poor clarity, the videos with frequent transitions between story segments, the videos with long, meaningless static shots, etc. Thus, to construct the high-quality dataset, we carefully filter out the low-quality videos before annotation.

2.3 Retrieval Uniqueness Guarantee

We ensure uniqueness through manual verification. Specifically, we set the candidate video pool size N_V (defined in Section 3.1 of the paper main body) as 200 for each nature language query. 199 randomly selected videos and the target video compose each candidate video pool. Annotators and verifiers are responsible for checking each candidate video pool, to guarantee the language-based retrieval will obtain the unique video result.

2.4 Dataset Statistics

In this section, the statistics of the Tex-COIN dataset are supplemented. All action classes are shown in Table 1.

2.5 Statistic Analysis Supplement

In this section, we supplement the statistic analysis in section 3.3. Specifically, from the statistic of the commonsense knowledge annotations (Figure 3(a)), it can be observed that each type of commonsense knowledge exhibits a long-tail distribution in terms of category quantity. This aligns with the common characteristic of naturally collected datasets [6, 14]. From the statistic of the sentence word count, it can be found that there are textual observations of different lengths, including longer sentences (containing more than 15 words) and shorter sentences (containing less than 9 words). From the perspective of text length, it can be observed that our dataset contains textual observation examples with varying amounts of information.

3 DATASET DOCUMENT SUPPLEMENT

Dataset Background As illustrated in Section 1, we propose a new dataset Tex-COIN as the testbed for the new task, COIN, to facilitate the development of hypothesis inference in the multi-modal domain.

Dataset Collection Our Tex-COIN dataset comes from two sources, the manually clipped TV shows and previous datasets, including the HC-STVG dataset [16], the ava-actions dataset [4], the Charades dataset [15] and the vidor dataset [14], as illustrated in Section 3.2. It is worth noticing that we do not use the labels of these datasets. The action labels from the ava-actions dataset [4] and the Charades dataset [15] are too sparse, whose videos are relabeled by us during the commonsense knowledge annotation process. The labels from other datasets are not consistent with our COIN task goal, which are not adopted. All videos are carefully selected from these sources to remove the videos unsuitable for the hypothesis inference annotation, like blurry videos. About 23% videos are saved for subsequent labeling, which leads to a heavy workload for data screening. In addition, we provide careful training and assessment to the screening personnel to ensure that they have no gender bias or other discrimination.

Dataset Labeling Each video is carefully labeled with commonsense knowledge and hypothesis-inference examples by the annotators. Some commonsense knowledge needs to be labeled frame by frame, like the scene and the sentiment. The frame rate is approximately 30 frames per second. Labeling all 30 frames of each second for the scene and the sentiment knowledge is not necessary. Thus, we annotate every 5th frame by extracting frames at equal intervals from one second. After labeling, we carefully check the annotations and remove the ones not agreed upon by all verifiers. We set the size of each video pool for each query as 150 and carefully check the videos in each video pool to prevent there being multiple correct videos corresponding to the given query.

Dataset Maintenance The dataset will be released follows the CC BY-NC-SA 4.0 protocol. We will hold and maintain this dataset for the long term. If there are any errors, we will update the dataset to guarantee its quality.

4 EXPERIMENT SUPPLEMENT

4.1 Implementation Detail

We implement our COINNet on a Linux server with Pytorch 1.4 and 8 Tesla V100 with 32GB memory. During the training process, we set the learning rate as $1e - 5$. The batch size is set to 8 and the optimizer is AdamW. In addition, we also employ some data augmentations, such as random horizontal flipping, random cropping, and so on. We adopt the same pretraining parameters for the COINNet as the Clip4Clip baseline [12].

Methods	CRET_C	CRET_F	Clip4Clip_C	Clip4Clip_F	COINNet
ACC	56.2	58.1	57.6	60.2	65.4

Table 2. The performance of all baselines on the **ACC** metric.

Methods	BM	BM+ δ_{Align}	BM+ δ_{Reason}	COINNet
ACC	59.6	62.1	64.7	65.4

Table 3. Ablation model performance on the **ACC** metric. "BM" represents the base model in the ablation study.

Methods	R@1	R@5	R@10	MdR	ACC	MoC
Cheetah	6.2	22.3	38.9	18	63.9	29.2
COINNet	7.3	25.3	39.9	16	65.4	31.9

Table 4. The performance comparison with the GMLLM state-of-the-art on the Tex-COIN dataset.

4.2 ACC Metric

We supplement the baseline comparison and the ablation study for the COINNet model with the ACC metric. The experiment results are shown in Table 2 and Table 3. It proves the effectiveness of the key modules, including the Knowledge-guided Cross-modal Alignment (δ_{Align}) and the Graph-based Non-parametric Reasoning (δ_{Reason}).

4.3 Comparison with GMLLM

We are interested in the performance of the Generative Multi-modal Large Language Model (GMLLM) on our COIN task. Thus, we choose the state-of-the-art, Cheetah[9], in the GMLLM domain, and evaluate its performance on the data annotation and the hypothesis inference prediction:

(1) Data Annotation. We randomly select 100 examples from our Tex-COIN dataset and re-label them with the Cheetah. 2 human verifiers from the top 50 universities in the QS World University Rankings in the world are employed to evaluate the quality of the Cheetah annotations. If both of them agree with the annotation, the annotation is considered qualified. However, the annotation accuracy of Cheetah is 61%, which does not satisfy our need for the high-quality dataset.

Methods	R@1	R@5	R@10	MdR	ACC	MoC
Base Model	5.7	21.5	34.9	20	59.6	27.9
$+\delta_{appearance}$	6.0	22.9	35.1	18	61.2	28.7
$+\delta_{clothing}$	5.9	22.3	35.4	19	61.4	28.9
$+\delta_{action}$	6.2	22.8	35.8	18	61.9	29.4
$+\delta_{sentiment}$	5.8	21.9	36.3	19	60.2	29.8
$+\delta_{scene}$	5.8	21.7	35.6	20	60.9	29.5
$+\delta_{Align}$	6.3	23.1	36.8	17	62.1	30.1

Table 5. Ablation study on the Video-CORE dataset about the commonsense knowledge. It is worth noting that each commonsense knowledge is individually added to the base model and tested. $+\delta_{Align}$ represents adding the guidance of all the commonsense knowledge together.

(2) Hypothesis Inference Prediction. We evaluate the Cheetah model on our Tex-COIN dataset, whose prediction accuracy for hypothesis-inference examples is shown in Table 4. Our COINNet performs better than the GMLLM state-of-the-art, Cheetah.

Possible reasons for the poor performance of the GMLLM are as follows: (1) Due to the limitation of training computing power, the generative multi-modal large language models at this stage need to be further developed for issues such as multiple images/videos [9, 17]. (2) The GMLLM is prone to being influenced by hallucinations [17], which highlights the necessity to further enhance their ability to deduce intricate logical connections.

4.4 Ablation Study

We add the guidance of each commonsense knowledge independently (including appearance, clothing, action, sentiment, and scene), to evaluate the effectiveness of each one. The experiment results are shown in Table 5. From it, we can observe that the model performs better after adding each of the commonsense knowledge guidance. It proves the reasonable design of the Knowledge-guided Cross-modal Alignment.

In addition, we are interested in the effect of the character’s physical description in the language-based retrieval process and design of the target experiment. Specifically, we only preserve the character’s physical description in each nature language query of our Tex-COIN dataset based on the widely used StanfordNLP tool and construct a new dataset, **Tex-COIN-M**. Then, we remove the action-flow reasoning module of our COINNet and evaluate the newly constructed model on the Tex-COIN-M dataset. Experiment results are shown in Table 6. From the table, it can be observed that the performance of the models on the Tex-COIN-M dataset drops sharply, compared with these models evaluated on the Tex-COIN dataset. It proves the models cannot effectively retrieve the correct target video from the video pool, only relying on the character description in the nature language query.

Methods	R@1	R@5	R@10	MdR
Clip4Clip	2.6	10.1	15.1	67
COINNet	3.1	12.5	19.2	56

Table 6. Model performance comparison on the Tex-COIN-M dataset.

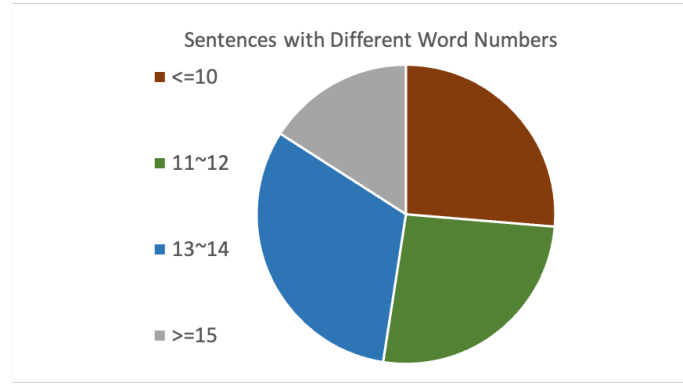


Fig. 2. The statistic of the sentences with different word numbers.

4.5 Case Study

We supplement the case study in Section 5.3. Specifically, we randomly select 7 examples from the Tex-COIN dataset, and show the experiment results in Figure 3. With the targeted designs, our COINNet makes the correct predictions.

5 TEXT-VIDEO RETRIEVAL TASK

We supplement the description of the text-video retrieval task in Section 1. Specifically, we add the text-video retrieval annotations for our Tex-COIN dataset to extend the dataset function, and experiment with the retrieval baselines on the Tex-COIN dataset. In this section, we will describe them in detail.

5.1 Dataset Annotation

Following the retroductive reasoning annotation, we provide the annotations for the text-video retrieval task on the Video-CORE dataset. Specifically, 4 trained annotators are responsible for the annotation. The process consists of two steps: (1) The annotators view all videos in the Tex-COIN dataset. Then, the key clips of the videos are cropped out and labeled with the language description. (2) 2 verifiers are responsible for conducting detailed checks on the annotations. If they all agree with the label, it is saved. Otherwise, it is re-labeled or discarded. After annotation, there are 9,983

Methods	R@1	R@5	R@10	MdR
Clip4Clip	5.2	19.5	32.7	21
Base Model	6.0	19.7	32.1	21
COINNet (Base Model + δ_{Align})	6.5	20.5	34.0	18

Table 7. Comparing our model, COINNet, with other baselines on the Tex-COIN dataset for the text-video retrieval task. δ_{Align} represents the Knowledge-guided Cross-modal Alignment design.

examples in our Tex-COIN dataset for the text-video retrieval task, which are split into train/val/test (9,033/200/750). The statistic of the text-video retrieval annotations is shown in Figure 2.

5.2 Experiments

We compare our COINNet model with the state-of-the-art on Tex-COIN dataset with the text-video retrieval annotations. In addition, we conduct the ablation study for our COINNet. The experiment results are shown in Table 7. From it, it can be found that our COINNet model performs best. We attribute the improvement to the effectiveness of the Knowledge-guided Cross-modal Alignment design.

6 MODEL SUPPLEMENT

In this section, we supplement the description of the COINNet model proposed by us in Section 4.

6.1 Knowledge-guided Cross-modal Alignment Supplement

We supplement the commonsense prediction (except the action prediction) described in Section 4.1.

(1) Sentiment and Scene Prediction. Considering that the sentiment knowledge and the scene knowledge are changed over time they are predicted frame by frame with the frame-level features \mathbf{F}_f . We use the i -th frame as an example, and then the prediction process for the i -th frame with the corresponding frame feature \mathbf{f}_f^i denoted as:

$$\mathbf{p}_{se}^i = \text{softmax}(MLP_{se}(\mathbf{f}_f^i)), \quad (1)$$

$$\mathbf{p}_{sc}^i = \text{softmax}(MLP_{sc}(\mathbf{f}_f^i)). \quad (2)$$

In them, MLP_{se} and MLP_{sc} are the MultiLayer Perceptron (**MLP**), which are applied to predict the sentiment and scene probabilities (\mathbf{p}_{sc}^i and \mathbf{p}_{se}^i), respectively.

(2) Appearance and Clothing Prediction. Regarding the appearance and clothing knowledge of the target person in the video, they are approximated as remaining constant throughout the video and are applied with the video-level query \mathbf{f}_v^2 to predict. In rare cases where there are special situations (such as costume changes), we require the model to predict the most relevant appearance and clothing knowledge for the textual observation description. We can represent the prediction process for the appearance and clothing knowledge as follows:

$$\mathbf{p}_{ap} = \text{softmax}(MLP_{ap}(\mathbf{f}_v^2)), \quad (3)$$

$$\mathbf{p}_{cl} = \text{softmax}(MLP_{cl}(\mathbf{f}_v^2)). \quad (4)$$

In the process, MLP_{ap} and MLP_{cl} are the MLP utilized to predict the appearance and clothing probabilities (\mathbf{p}_{ap} and \mathbf{p}_{cl}), respectively.

The model learns to make all commonsense knowledge predictions simultaneously. To achieve this, we use the cross-entropy loss function to calculate the losses for all commonsense knowledge, including l_{se} (sentiment), l_{sc} (scene), l_{ac} (action), l_{ap} (appearance), and l_{cl} (clothing). To obtain the complete knowledge-guided loss l_{total} , we sum up all the individual losses mentioned above:

$$l_{total} = l_{ac} + l_{se} + l_{sc} + l_{ap} + l_{cl}. \quad (5)$$

7 APPLICATION DISCUSSION

To facilitate multi-modal hypothesis inference, we propose the COIN task, the task-corresponded dataset, Tex-COIN, and the strong baseline, COINNet. We think one important actual application of our work, is to assist in building a human-like inference AI system. Because the hypothesis inference ability is important for human in the daily life, as detailed in the lines 29-32 of our paper. Similarly, for the human-like inference AI system, this hypothesis-inference ability provides the capability of tracing the causes of phenomena, which can be applied to practical scenarios. One application scenario is reading comprehension. Specifically, in the context of reading comprehension, based on textual descriptions, the hypothesis-inference AI system can establish connections between the text content and real-life visual memories, thereby truly understanding the causes behind the textual descriptions. Then, it can better accomplish subsequent tasks such as question answering according to the text. In addition, another significant application of COIN is in the domain of intelligent security. According to the description of the witness, AI system with the hypothesis-inference capability can search for evidence and infer the criminal process from a massive amount of video footage.

REFERENCES

- [1] Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. 2020. Fine-grained video-text retrieval with hierarchical graph reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10638–10647.
- [2] Sheng Fang, Shuhui Wang, Junbao Zhuo, Qingming Huang, Bin Ma, Xiaoming Wei, and Xiaolin Wei. 2022. Concept propagation via attentional knowledge graph reasoning for video-text retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*. 4789–4800.
- [3] Zerun Feng, Zhimin Zeng, Caili Guo, and Zheng Li. 2020. Exploiting visual semantic reasoning for video-text retrieval. *arXiv preprint arXiv:2006.08889* (2020).
- [4] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. 2018. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6047–6056.
- [5] Yu-Jung Heo, Eun-Sol Kim, Woo Suk Choi, and Byoung-Tak Zhang. 2022. Hypergraph Transformer: Weakly-Supervised Multi-hop Reasoning for Knowledge-based Visual Question Answering. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 373–390.
- [6] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. 2020. Action genome: Actions as compositions of spatio-temporal scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10236–10247.
- [7] Dazhi Jiang, Hao Liu, Geng Tu, and Runguo Wei. 2023. Window transformer for dialogue document: a joint framework for causal emotion entailment. *International Journal of Machine Learning and Cybernetics* (2023), 1–11.
- [8] Jiangnan Li, Fandong Meng, Zheng Lin, Rui Liu, Peng Fu, Yanan Cao, Weiping Wang, and Jie Zhou. 2022. Neutral utterances are also causes: Enhancing conversational causal emotion entailment with social commonsense knowledge. *arXiv preprint arXiv:2205.00759* (2022).
- [9] Juncheng Li, Kaihang Pan, Zhiqi Ge, Minghe Gao, Hanwang Zhang, Wei Ji, Wenqiao Zhang, Tat-Seng Chua, Siliang Tang, and Yueting Zhuang. 2024. Fine-tuning Multimodal LLMs to Follow Zero-shot Demonstrative Instructions. In *The Twelfth International Conference on Learning Representations*.
- [10] Mengze Li, Tianbao Wang, Jiahe Xu, Kairong Han, Shengyu Zhang, Zhou Zhao, Jiaxu Miao, Wenqiao Zhang, Shiliang Pu, and Fei Wu. 2023. Multi-modal Action Chain Abductive Reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 4617–4628.

The short hair woman in a blouse is throwing the food waste into the trash can.

↓ Hypothesis Inference

Retrieved Video:

COINNet: (✓)



Action Flow

COINNet: ⇨ eat ⇨ clean ⇨ pick up ⇨ hold ⇨ (✓)

Commonsense Knowledge Prediction COINNet: (✓)

Appearance: woman, short hair, adult;
Clothing: long sleeve red, yellow, and gray blouse, no hat, no glasses, no tie;
Action: cook; **Sentiment:** neutral; **scene:** kitchen.

Fig. 3. 7 examples of our COINNet model predictions.

- [11] Chen Liang, Wenguan Wang, Tianfei Zhou, and Yi Yang. 2022. Visual abductive reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15565–15575.
- [12] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. 2022. CLIP4Clip: An empirical study of CLIP for end to end video clip retrieval and captioning. *Neurocomputing* 508 (2022), 293–304.
- [13] Kanagasabai Rajaraman, Saravanan Rajamanickam, and Wei Shi. 2023. Investigating Transformer-Guided Chaining for Interpretable Natural Logic Reasoning. In *Findings of the Association for Computational Linguistics: ACL 2023*. 9240–9253.
- [14] Xindi Shang, Donglin Di, Junbin Xiao, Yu Cao, Xun Yang, and Tat-Seng Chua. 2019. Annotating Objects and Relations in User-Generated Videos. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*. ACM, 279–287.
- [15] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer, 510–526.
- [16] Zongheng Tang, Yue Liao, Si Liu, Guanbin Li, Xiaojie Jin, Hongxu Jiang, Qian Yu, and Dong Xu. 2021. Human-centric spatio-temporal video grounding with visual transformers. *IEEE Transactions on Circuits and Systems for Video Technology* 32, 12 (2021), 8238–8249.
- [17] Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. 2024. MMICL: Empowering Vision-language Model with Multi-Modal In-Context Learning. In *The Twelfth International Conference on Learning Representations*.
- [18] Weixiang Zhao, Yanyan Zhao, Zhuojun Li, and Bing Qin. 2023. Knowledge-bridged causal interaction network for causal emotion entailment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 14020–14028.
- [19] Daoming Zong and Shiliang Sun. 2023. MCOMET: Multimodal Fusion Transformer for Physical Audiovisual Commonsense Reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 6621–6629.

Observation: A man with a black pant is wiping his face with a towel.

⇩ Hypothesis Inference

Retrieved Video:

COINNet: (✓)



Action Flow

COINNet: ⇨ pick up ⇨ hold ⇨ (✓)

Commonsense Knowledge Prediction COINNet: (✓)

Appearance: man, short hair, adult;

Clothing: short sleeve gray T-shirt, long black pant, no hat, no glasses, no tie;

Action: wash; **Sentiment:** neutral; **scene:** bathroom.

Observation: There is an short hair adult in a red T-shirt going to the hospital.

⇩ Hypothesis Inference

Retrieved Video:

COINNet: (✓)



Action Flow

COINNet: ⇨ stand up ⇨ walk out ⇨ (✓)

Commonsense Knowledge Prediction COINNet: (✓)

Appearance: man, short hair, adult;

Clothing: short sleeve red T-shirt, long black pant, no hat, no glasses, no tie;

Action: fall down; **Sentiment:** neutral; **scene:** stairs.

Observation: *The long hair woman in a gray and black hoodie is drying clothes.*

⇩ **Hypothesis Inference**

Retrieved Video:

COINNet: (✓)



Action Flow

COINNet: ⇨ wring out ⇨ (✓)

Commonsense Knowledge Prediction COINNet: (✓)

Appearance: woman, long hair, adult;

Clothing: long sleeve black and gray hoodie, long black pant, no hat, no glasses, no tie; **Action:** wash; **Sentiment:** neutral; **scene:** laundry room.

Observation: *The man with a black T-shirt is cleaning the mud on the shoe.*

⇩ **Hypothesis Inference**

Retrieved Video:

COINNet: (✓)



Action Flow

COINNet: ⇨ bend over ⇨ pick up ⇨ hold ⇨ stand up ⇨ (✓)

Commonsense Knowledge Prediction COINNet: (✓)

Appearance: man, short hair, adult;

Clothing: short sleeve black T-shirt, black pant, no hat, no glasses, no tie; **Action:** throw; **Sentiment:** neutral; **scene:** bathroom.

Observation: *There is an short hair adult with a shirt going to the hospital.*

⇩ **Hypothesis Inference**

Retrieved Video:

COINNet: (✓)



Action Flow

COINNet: ⇨ walk out ⇨ (✓)

Commonsense Knowledge Prediction COINNet: (✓)

Appearance: man, short hair, adult;
Clothing: long sleeve yellow, gray, and white shirt, brown pant, no hat, no glasses, no tie; **Action:** sneeze; **Sentiment:** neutral; **scene:** kitchen.

Observation: *The adult in a black T-shirt is wiping his nose with a paper.*

⇩ **Hypothesis Inference**

Retrieved Video:

COINNet: (✓)



Action Flow

COINNet: ⇨ pick up ⇨ hold ⇨ (✓)

Commonsense Knowledge Prediction COINNet: (✓)

Appearance: man, short hair, adult;
Clothing: short sleeve black T-shirt, long pant, no hat, no glasses, no tie;
Action: sneeze; **Sentiment:** neutral; **scene:** bedroom.