Multimodal Label Relevance Ranking via Reinforcement Learning

Taian Guo*, Taolin Zhang*, Haoqian Wu, Hanjun Li, Ruizhi Qiao, Xing Sun

Tencent Youtu Lab

Tsinghua Shenzhen International Graduate School, Tsinghua University

Outline

- Motivation
- Problem Setting
- Proposed Method
- Proposed Dataset
- Experiment Results
- Main Contributions

Label Confidence vs. Label Relevance

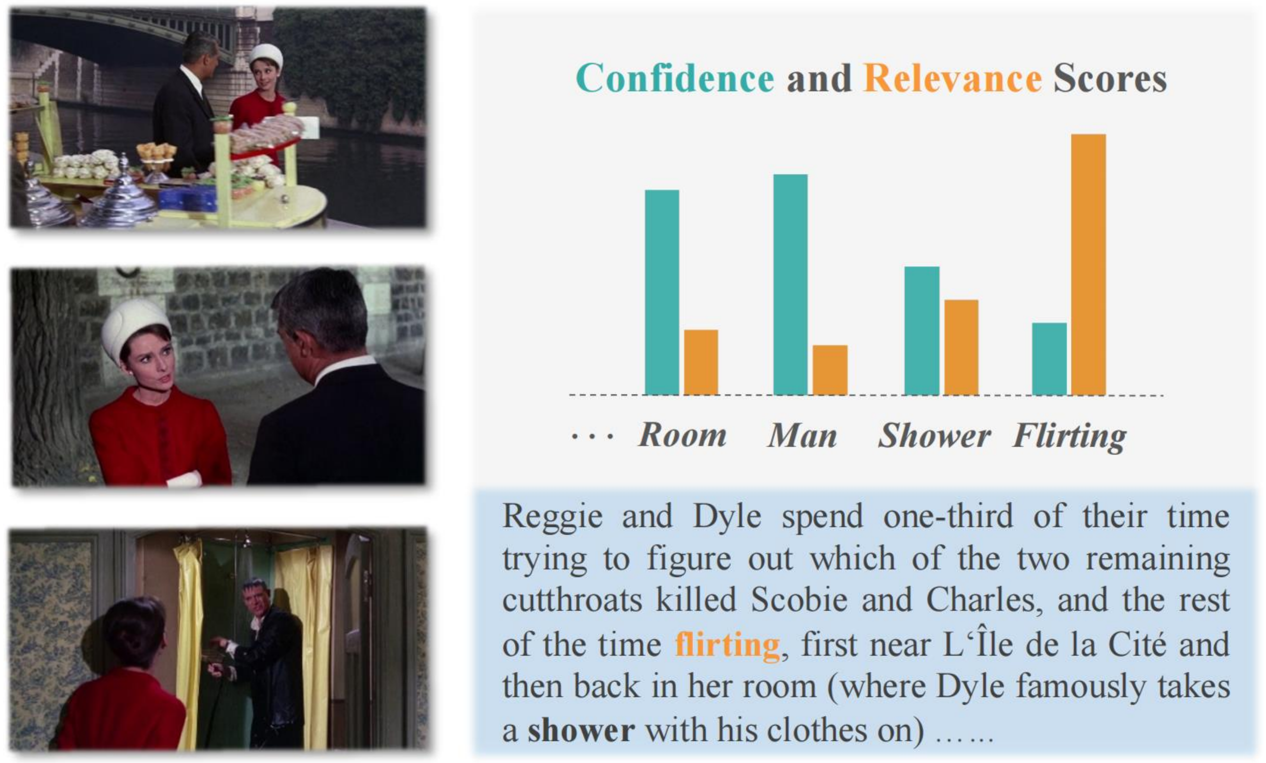


Fig. 1: Illustration of the Difference between Label Confidence and Label Relevance. This figure provides an example of a movie footage consisting of three consecutive keyframes and its scene description. Generally, conventional label confidence tends to place more emphasis on the tangible objects, whereas the proposed label relevance better reveals the relations between labels and the real scene which they correspond to.

Definitions

- **Definition 1 (Label Confidence).** *Given a multi-label classification task with a set of labels $\mathcal{L} = \{l_1, l_2, \ldots, l_n\}$, an instance x is associated with a label subset $\mathcal{L}_x \subseteq \mathcal{L}$. The label confidence of a label l_i for instance x, denoted as $C(l_i|x)$, is defined as the* ***probability*** *that l_i is a* ***correct*** *label for x,* i.e., $C(l_i|x) = P(l_i \in \mathcal{L}_x|x)$. (1)
- **Definition 2 (Label Relevance).** *The label relevance of a label l_i for instance x, denoted as $R(l_i|x)$, is defined as the* ***degree*** *of* ***association*** *between l_i and x,* i.e., $R(l_i|x) = f(l_i, x)$, (2) *where f is a function that measures the degree of association between l_i and x.*

The Importance of Label Relevance

- Label confidence typically refers to the estimation from a model about the probability of a label's occurrence, while label relevance primarily denotes the significance of the label to the **primary theme** of multimodal inputs.
- Relevance labels bear a closer alignment with human preferences.
- Ranking the labels in order of relevance can be employed to emphasize the important labels.

Multimodal Label Relevance Ranking

Problem Setting

- Given V video clips, where the j-th clip consists of frames $F^j = [F_0^j, F_1^j, ..., F_{N-1}^j]$, with N representing the total number of frames extracted from a video clip, and j ranging from 0 to $V-1$.
- Each video clip is accompanied by text descriptions T^j and a set of recognized labels denoted as $\mathcal{L}^j$, where $\mathcal{L}^j = \{l_0^j, l_1^j, ..., l_i^j, ..., l_{|\mathcal{L}^j|-1}^j\}$, and $|L^j|$ is the number of labels in the j-th video clip.
- The objective of label relevance ranking is to learn a ranking function $f_{\text{rank}} : F^j, T^j, L^j \rightarrow U^j$, where $U^j = [u_0^j, u_1^j, ..., u_i^j, ..., u_{|\mathcal{L}^j|-1}^j]$ represents the ranking result of the label set L^j.

Metrics

- NDCG: Normalized Discounted Cumulative Gain.
- NDCG@k : For each video clip, we compute NDCG@k for the top k labels.

Overall Framework of LR^{2}PPO

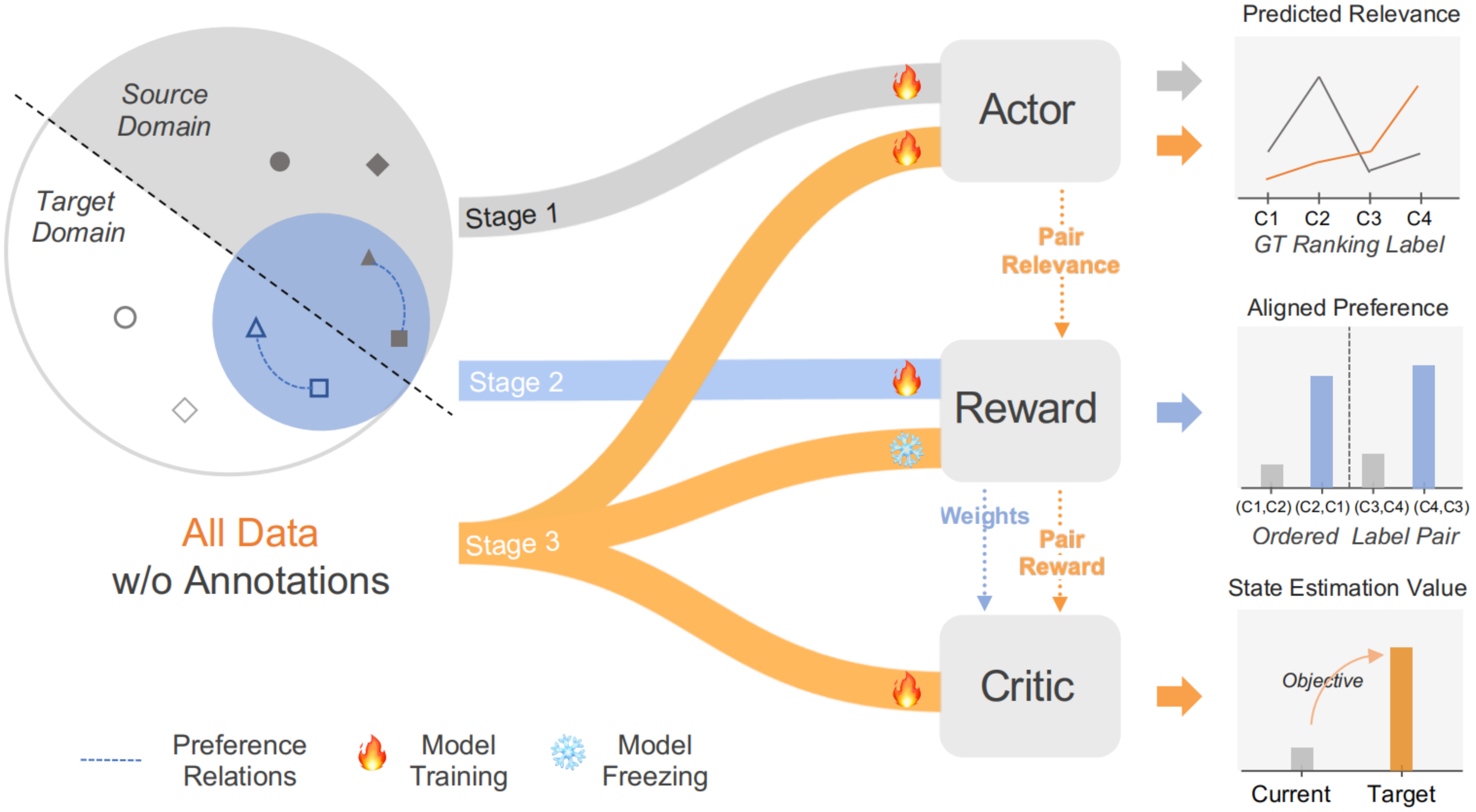


Fig. 2: Illustration of the training paradigm of LR2PPO. Each stage takes multimodal data as input but differs in terms of specific data division and annotation type. Technically, in Stage 1, data from the source domain is employed to establish a label relevance ranking base model (i.e., ***Actor***). Stage 2 involves preference data to train a ***Reward*** model. Finally, in Stage 3, ***Critic*** model interacts with the first two models and all data w/o annotations is utilized to boost the performance of the ***Actor***, which will solely be applied in the inference stage.

Stage 1 and Stage 2 of LR^{2}PPO Framework

Stage 1. Label Relevance Ranking Base Model.

- During Stage 1, the training of the label relevance ranking base model adopts a supervised paradigm, i.e., it is trained on the source domain based on manually annotated relevance categories (high, medium and low). SmoothL1Loss is calculated for optimization:

$$L_{\text{SmoothL1}}(p) = \begin{cases} 0.5(p-y)^2/\beta & \text{if } |p-y| < \beta \\ |p-y| - 0.5\beta & \text{otherwise,} \end{cases}$$

Stage 2. Reward Model.

- We train a reward model on the target domain in stage 2. With a few label pair annotations on the target domain, along with augmented pairs sampled from the source domain, the reward model can be trained to assign rewards to the partial order relationships between label pairs of a given clip. This kind of partial order relation annotation aligns with human preference of label relevance ranking, thus benefiting relevance ranking performance with limited annotation data. The loss function adopted for the training:

$$L_{RM}(g_{ini}, g_c) = \max(0, m_R - (R([g_{ini}, g_c]) - R([g_{ini}, \text{flip}(g_c)]))),$$

Stage 3 of LR²PPO Framework

Stage 3. LR²PPO. State Definition and More

- **State s_t:** the order of a group of labels (specifically, a label pair) at timestep t
- **Action a_t:** the policy network (aka. actor model) predicts the relevance score of the labels and ranks them from high to low to obtain a new label order as next state s_{t+1}, which is considered a state transition, or action a_t
- **Policy π_θ:** the forementioned process of state transition
- **Reward r_t:** obtained by the reward model with state s_t and action a_t as inputs

Stage 3. LR²PPO. Policy Loss Definition and More

- **Representation for the Change in Label Order:** complete probability vector, i.e., state transition, instead of the maximum component
- **Partial Order Function Definition:** $H_{partial}(p_t^1, p_t^2) = \max(0, m - (p_t^1 - p_t^2)),$
- **Partial Order Ratio $r_t'(\theta)$ as Adjustment for Advantage:**
- **Policy Function Loss:** $L^{PF}_{LR^2PPO}(\theta) = -\mathbb{E}_t\left(r_t'(\theta)abs(\hat{A}_t)\right).$

$$r_t'(\theta) = \begin{cases} -H^{partial}(p_t^1, p_t^2) & \hat{A}_t \geq \delta \\ -H^{partial}(p_t^2, p_t^1) & \hat{A}_t < \delta. \end{cases}$$

Procedure of LR^2PPO Core Algorithm

Algorithm 1 Label Relevance Ranking with Proximal Policy Optimization (LR^2PPO), Actor-Critic Style

Input: Policy network $\pi_{\theta_{\text{old}}}$, state value network $V_{\omega_{\text{old}}}$, number of timesteps T, number of trajectories in an iteration N_{Trajs}, number of epochs K, minibatch size M

Output: Policy network parameter θ, state value network parameter ω

```
Initialization:
Initialize θ_old and ω_old with base model and reward model
LOOP Process
for iteration = 1, 2, ... do
    for n_traj = 1, 2, ..., N_Trajs do
        Run policy π_θold and state value network V_ωold in environment for T timesteps
        Compute advantage estimates Â_1, ..., Â_T according to Eq. (6)
    end for
    Compute joint loss L_LR²PPO according to Eq. (11)
    Optimize surrogate L_LR²PPO with respect to θ and ω, with K epochs and mini-
    batch size M ≤ N_Trajs · T
    θ_old ← θ, ω_old ← ω
end for
return: θ, ω
```

The pseudo-code of our LR^2PPO is provided in Algorithm 1.

LRMovieNet Dataset

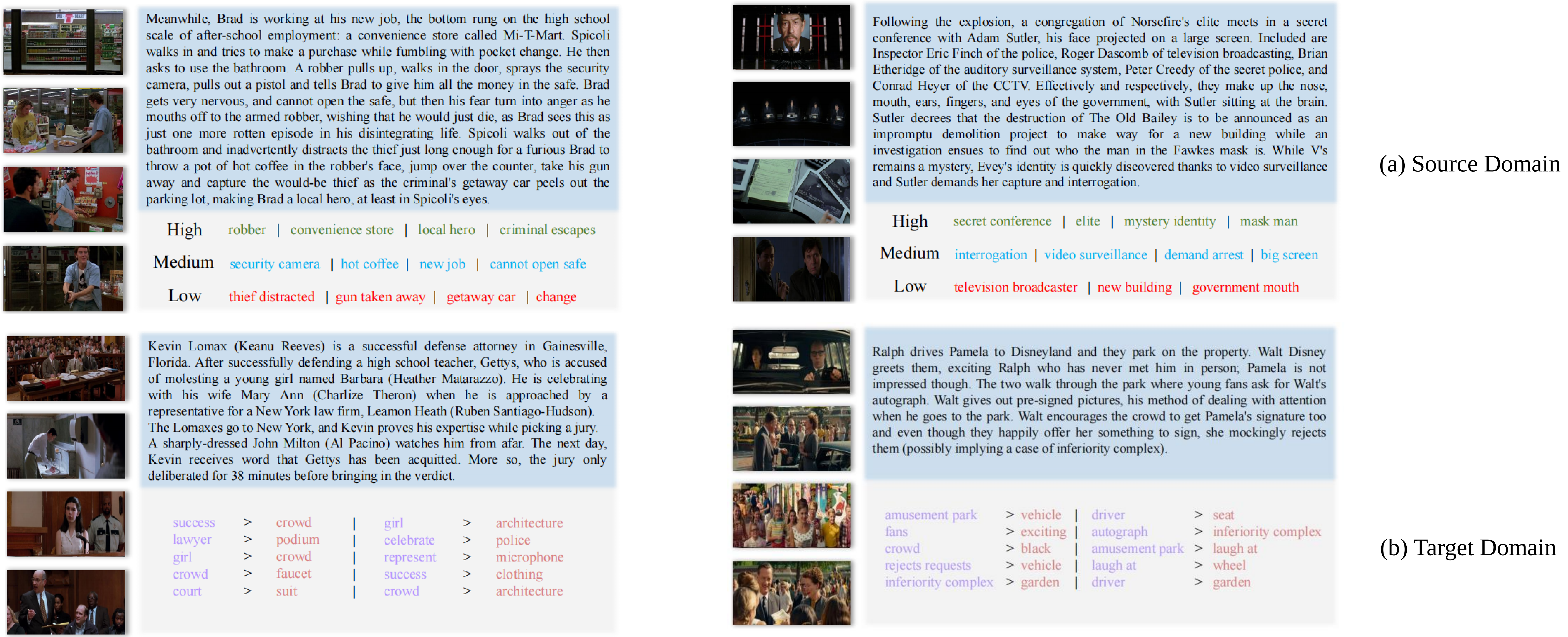
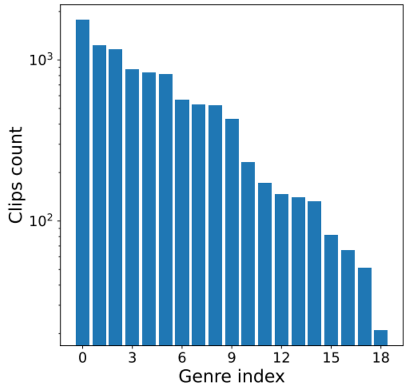
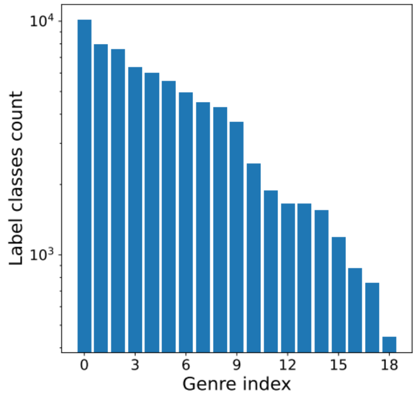


Fig. C.2: Annotated training samples in source and target domains. The red, blue and green labels listed in the upper subfigure represent low, medium and high in ground truth in the source domain, respectively. For each label pair in the lower subfigure, the left label are more relevant than the right in accordance with the video episode context (*i.e.*, descriptions and frames). Best viewed in color and zoomed in.

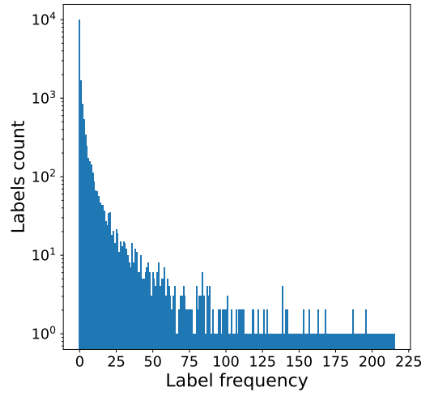
LRMovieNet Dataset



(a) Clips count in different genres



(b) Label classes count in different genres



(c) Labels count about label frequency

Fig. C.1: Data statistics of LRMovieNet.

Genre Index	Genre Name	Clips Count
0	Drama	1765
1	Action	1224
2	Thriller	1154
3	Sci-Fi	869
4	Crime	829
5	Adventure	814
6	Comedy	562
7	Mystery	525
8	Fantasy	520
9	Romance	427
10	Biography	232
11	War	171
12	Horror	147
13	Family	140
14	History	132
15	Music	82
16	Western	66
17	Sport	51
18	Musical	21

(a) Details about clips count in different genres

Genre Index	Genre Name	Classes Count
0	Drama	10088
1	Action	7923
2	Thriller	7573
3	Sci-Fi	6315
4	Adventure	5999
5	Crime	5542
6	Comedy	4933
7	Fantasy	4468
8	Mystery	4277
9	Romance	3688
10	Biography	2443
11	War	1879
12	Horror	1658
13	History	1652
14	Family	1543
15	Music	1191
16	Western	872
17	Sport	757
18	Musical	446

(b) Details about label classes count in different genres

Table C.1: Details about number of video clips and label classes in all videos of different genres in LRMovieNet.

Results on LRMovieNet Dataset

Method		NDCG @ 1	NDCG@3	NDCG@5	NDCG@10	NDCG@20
OV-based	CLIP [47]	0.5523	0.5209	0.5271	0.6009	0.7612
	MKT [23]	0.3517	0.3533	0.3765	0.4704	0.6774
LTR-based	PRM [45]	0.6320	0.6037	0.6083	0.6650	0.8022
	DLCM [1]	0.6153	0.5807	0.5811	0.6310	0.7866
	ListNet [9]	0.5947	0.5733	0.5787	0.6438	0.7872
	GSF [2]	0.594	0.571	0.579	0.643	0.787
	SetRank [44]	0.6337	0.6038	0.6125	0.6658	0.8030
	RankFormer [8]	0.6350	0.6048	0.6108	0.6655	0.8033
Ours	LR^2PPO (S1)	0.6330	0.6018	0.6061	0.6667	0.8021
	LR^2PPO	**0.6820**	**0.6714**	**0.6869**	**0.7628**	**0.8475**

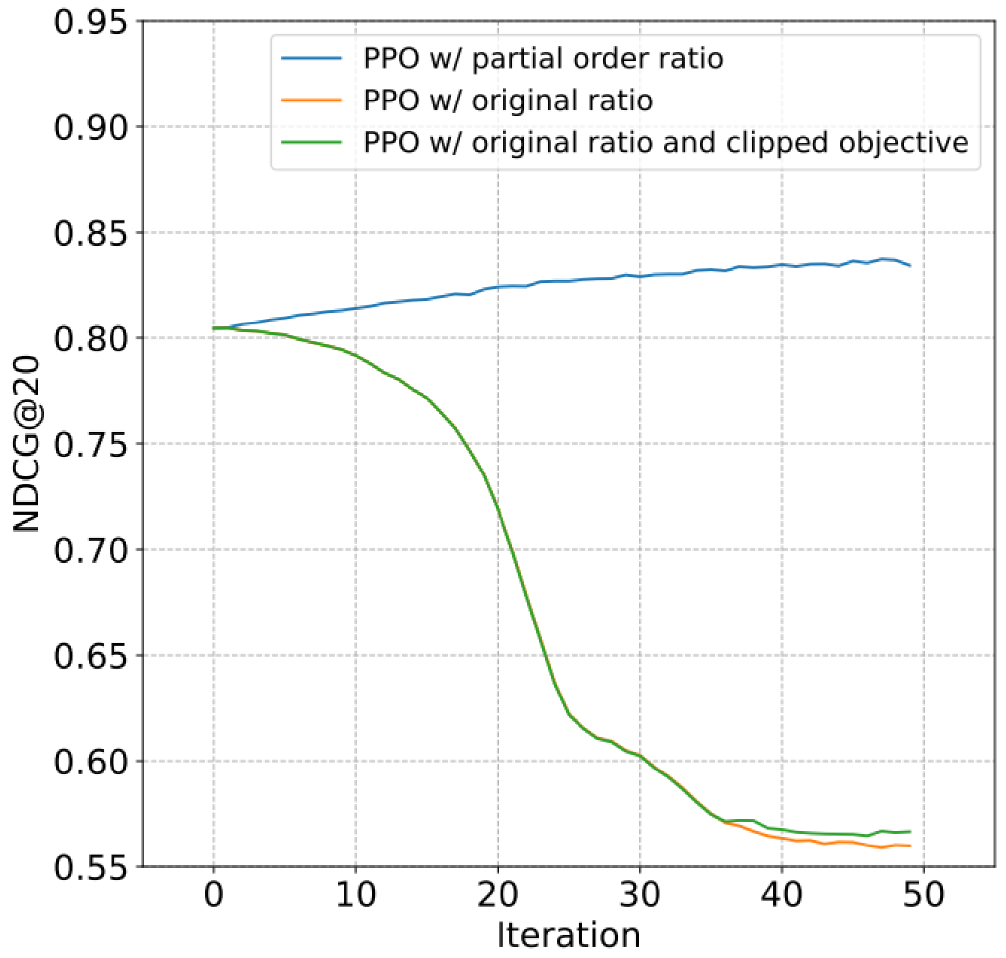
Table 1: State-of-the-art comparison for Label Relevance Ranking task on the LRMovieNet dataset. **Bold** indicates the best score.

Results on MSLR-Web10K → MQ2008

Method	NDCG @ 1	NDCG@3	NDCG@5	NDCG@10	NDCG@20
PRM [45]	0.5726	0.5804	0.5973	0.6407	0.7603
DLCM [1]	0.5983	0.6025	0.6125	0.6797	0.7744
ListNet [9]	0.5449	0.5575	0.5699	0.6324	0.7467
GSF [2]	0.6004	0.6265	0.6471	0.7054	0.7892
SetRank [44]	0.5299	0.5380	0.5555	0.6083	0.7365
RankFormer [8]	0.5684	0.5511	0.5643	0.6164	0.7458
LR^{2}PPO	**0.6496**	**0.6830**	**0.7033**	**0.7710**	**0.8240**

Table 2: State-of-the-art comparison on traditional datasets for label relevance ranking on the MSLR-Web10K → MQ2008 transfering task.

Influence of Key Designs of LR^{2}PPO



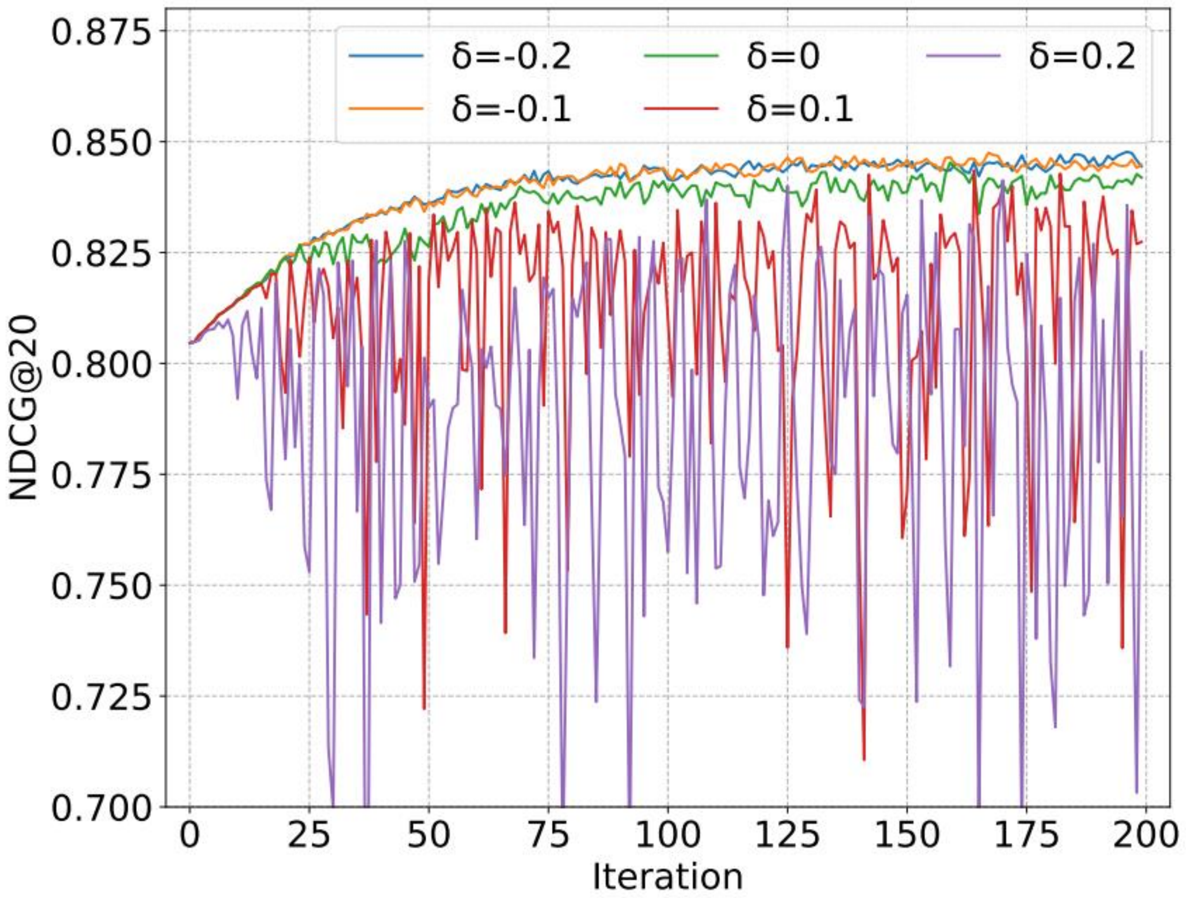


Fig. 3: NDCG curves during training. (a) PPO with different ratio design. Original ratio in PPO is not applicable to the definitions of state and action in the ranking task, leading to a training collapse, while our proposed partial order ratio solves this problem. (b) PPO with different thresholds δ in $r_t'(\theta)$. A small negative threshold $\delta = -0.1$ stabilizes the training, leading to superior performance.

Influence of Annotation Proportion in Target Domain

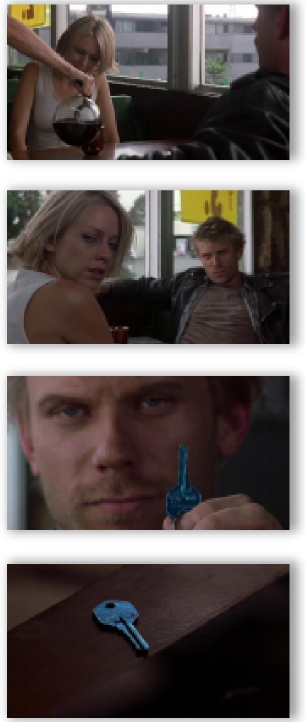
Annotation Proportion	Reward Model Accuracy	NDCG@1	NDCG@3	NDCG@5	NDCG@10	NDCG@20
0%	-	0.6330	0.6018	0.6061	0.6667	0.8021
5%	0.7697	0.6787	0.6581	0.6770	0.7514	0.8416
10%	0.7757	0.6820	0.6714	0.6869	0.7628	0.8475
20%	0.7837	0.6800	0.6784	0.6980	0.7667	0.8506
40%	0.7866	0.6830	0.6682	0.6877	0.7617	0.8467

Table 3: Stage 2 and 3 results with different annotation proportions in target domain.

Qualitative Assessment

Wladek goes to the emergency address he was given, where he surprisingly meets Dorota, who is now married, pregnant, and her brother dead. Dorota and her husband hide Wladek in another vacant apartment, where there is a piano, but his new caretaker, Szalas, is very slack about smuggling in food, and Wladyslaw once more faces starvation, and at one point almost dies of jaundice. Dorota and her husband visit him, finding him gravely ill. They report that Szalas had been collecting money from generous and unwitting donors and had pocketed it all, leaving Wladek to die in isolation.

CLIP	apartment	visit	paper	room	movie
NDCG@5: 0.65	0.982	0.973	0.973	0.967	0.961
PRM	to raise funds	seriously ill	slack off	caretaker	apartment
NDCG@5 = 0.74	1.79	1.75	1.60	1.59	1.55
LR^2PPO	smuggling in food	seriously ill	slack off	husband	caretaker
NDCG@5 = 0.84	14.0	12.2	10.8	10.6	10.5



Enraged, Diane hired Joe, a hit-man, to kill Camilla. Diane paid Joe the cash and showed him Camilla's headshot at Winkie's, where a waitress named Betty served them and a customer Diane dreamt as Dan watched them arrange the hit. Joe tells Diane that once he has completed the job, he will leave a blue key in her apartment -- the exact key that Diane has now found.

CLIP	coffee	apartment	hair	refer to	coffee shop
NDCG@5: 0.43	0.989	0.985	0.985	0.984	0.979
PRM	customer	coffee shop	apartment	enraged	dining room
NDCG@5 = 0.53	1.63	1.24	1.18	1.01	0.910
LR^2PPO	apartment	cash	customer	enraged	key
NDCG@5 = 0.91	11.5	10.6	10.6	10.4	10.1

Fig.4: Comparison between LR^2PPO and other state-of-the-art ranking methods. The red, blue and green labels listed after the method represent low, medium and high in ground truth, respectively. The value below each label represents the corresponding relevance score.

Main Contributions

- We recognize the significant role of label relevance, and analyze the limitations of previous ranking methods when dealing with label relevance. To solve this problem, we propose a multimodal label relevance ranking approach to rank the labels according to the relevance between label and the multimodal input. This the first work to explore the ranking in the perspective of label relevance.

- To better generalize the capability to new scenarios, we design a paradigm that transfers label relevance ranking ability from the source domain to the target domain. Besides, we propose the LR^2PPO (Label Relevance Ranking with Proximal Policy Optimization) to effectively mine the partial order relations among labels.

- To better evaluate the effectiveness of LR^2PPO, we annotate each video clip with corresponding class labels and their relevance order of the MovieNet dataset, and develop a new multimodal label relevance ranking bench-mark dataset, LRMovieNet (Label Relevance of MovieNet). Comprehensive experiments on this dataset and traditional LTR datasets demonstrate the effectiveness of our proposed LR^2PPO algorithm.

Thank you for listening

https://arxiv.org/abs/2407.13221

https://github.com/ChazzyGordon/LR2PPO