# Ranking Distillation for Open-Ended Video Question Answering with Insufficient Labels

Tianming Liang[1], Chaolei Tan[1], Beihao Xia[2], Wei-Shi Zheng[1], Jian-Fang Hu[1*]
[1]Sun Yat-sen University, China
[2]Huazhong University of Science and Technology, China
{liangtm, tanchlei}@mail2.sysu.edu.cn, hujf5@mail.sysu.edu.cn

## Abstract

*This paper focuses on open-ended video question answering, which aims to find the correct answers from a large answer set in response to a video-related question. This is essentially a multi-label classification task, since a question may have multiple answers. However, due to annotation costs, the labels in existing benchmarks are always extremely insufficient, typically one answer per question. As a result, existing works tend to directly treat all the unlabeled answers as negative labels, leading to limited ability for generalization. In this work, we introduce a simple yet effective ranking distillation framework (RADI) to mitigate this problem without additional manual annotation. RADI employs a teacher model trained with incomplete labels to generate rankings for potential answers, which contain rich knowledge about label priority as well as label-associated visual cues, thereby enriching the insufficient labeling information. To avoid overconfidence in the imperfect teacher model, we further present two robust and parameter-free ranking distillation approaches: a pairwise approach which introduces adaptive soft margins to dynamically refine the optimization constraints on various pairwise rankings, and a listwise approach which adopts sampling-based partial listwise learning to resist the bias in teacher ranking. Extensive experiments on five popular benchmarks consistently show that both our pairwise and listwise RADIs outperform state-of-the-art methods. Further analysis demonstrates the effectiveness of our methods on the insufficient labeling problem.*

## 1. Introduction

Video question answering [23, 35, 65] is one of the most popular research domains to explore the capability of AI models in understanding videos and languages. In this work, we rethink a fundamental task within this domain — **open-ended video question answering (OE-VQA)** [20, 50, 52, 59], which requires the model to find the correct answers in a large vocabulary (with over 1K possible answers) in response to the question regarding a given video.

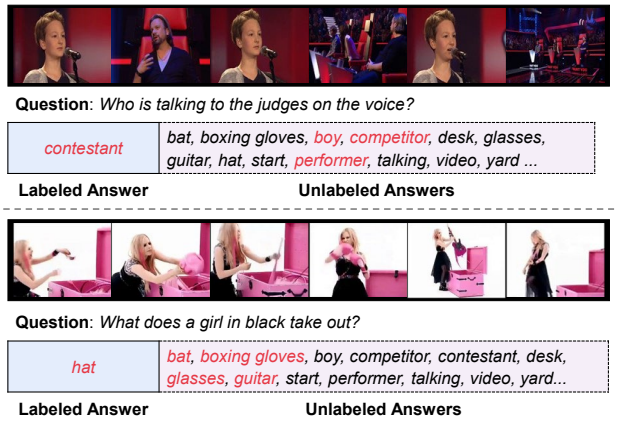


Figure 1. Two examples about the insufficient labeling problem in MSRVTT-QA dataset. The correct answers to each questions are colored in *red*. Existing OE-VQA methods tend to directly regard the entire unlabeled set as negative answers.

Essentially, OE-VQA is supposed to be a multi-label classification task, as there is always more than one answer corresponding to a question. However, due to the cost of annotation, the labels provided by existing public benchmarks are extremely insufficient, typically one labeled answer per question, as illustrated in Figure 1. In this study, we term this challenge as an **insufficient labeling problem** in OE-VQA. This problem is critical but surprisingly neglected by even state-of-the-art approaches [27, 45, 49, 53], which tend to formulate OE-VQA as a regular multi-class classification task by directly recognizing all the unlabeled answers as negative labels, as shown in Figure 2(a). These approaches ignore the numerous potentially correct answers, leading to a limited ability in open-world video question answering. To overcome this problem without extra manual labeling, three baseline schemes could be preliminarily considered:

**(i)** *Label smoothing* prevents overfitting the hard labels by replacing them with smoothed ones, as shown in Fig-

*Corresponding author.

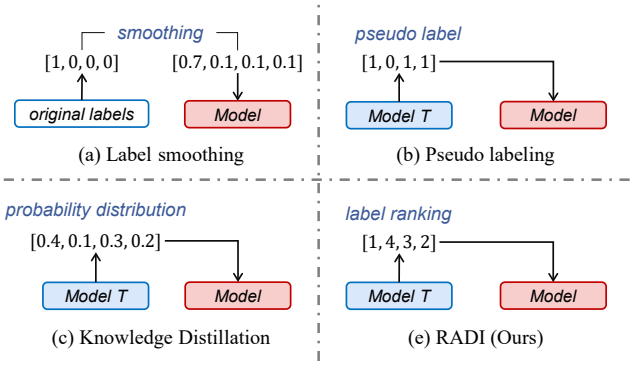


Figure 2. Comparison between different potential schemes for the insufficient labeling problem, where *Model T* serves as a teacher model to enrich the label information.

ure 2(a). However, it treats all negative labels equally, failing to offer further hints about the potential answers.

**(ii)** *Pseudo labeling* aims to generate pseudo labels with a trained model for label augmentation, as shown in Figure 2(b). However, the accumulation of false positive labels during this process leads to incremental noise, thus limiting further improvements in performance.

**(iii)** *Distribution distillation* aims to utilize the probability distributions of a trained teacher model to provide additional soft label information, as shown in Figure 2(c). However, in the context of OE-VQA, only incomplete labels are available for training the teacher model. Distribution distillation is sensitive to the noise and bias in the imperfect teacher model, which limits its potential performance.

This work presents a **RAnking DIstillation framework** (**RADI**), which further mitigates the insufficient labeling problem by overcoming the limitations of the above baselines. As shown in Figure 2(d), RADI utilizes answer rankings from a teacher model, which is trained with incomplete labels, to enrich the label information for training the student model. Indeed, RADI can be regarded as a slack form of distribution distillation, since it relaxes the matching constraint from probability distribution to the label ranking. In contrast to label smoothing and pseudo labeling, RADI can provide rich inter-label information while avoiding potential risks of hard pseudo labels. This is particularly beneficial for OE-VQA that typically involves a large label set. Compared with distribution distillation, RADI is more robust to the insufficient labeling problem, since it depends on the relative ordering of answers instead of absolute scores.

RADI is flexible and can be integrated with existing off-the-shelf learning-to-rank (LTR) methods. However, regular LTR methods might not be adequate to address the extremely insufficient labeling problem, since they are sensitive to the ranking position of each answer. Indeed, strictly aligning the rankings between the teacher and student models may lead to convergence similar to that of distribution distillation. Therefore, to further enhance the robustness of RADI, we design two alternate distillation approaches—*pairwise ranking distillation* and *listwise ranking distillation*. In general, the pairwise approach tends to learn the relative priority by pairwise comparison, while the listwise approach aims to directly learn the absolute orders of a ranked list. In the pairwise ranking distillation, we introduce soft margins to adaptively relax the optimization constraints to the uncertain pairwise rankings. In the listwise ranking distillation, we design two ranking-based sampling strategies to enable distillation on partial list. We empirically demonstrate that the two approaches can effectively avoid overconfidence in the teacher's noisy rankings. Moreover, both approaches are parameter-free, and bring no additional burden at inference. We summarize our main contributions below:

- We reveal the insufficient labeling problem in OE-VQA, and present an effective ranking distillation framework RADI to overcome it without extra manual annotation.
- We design two robust distillation strategies to further enhance the robustness of RADI to noisy rankings.
- We conduct extensive experiments on five popular OE-VQA benchmarks to demonstrate the improvement of RADI over state-of-the-art (SOTA) models and the effectiveness of the two distillation strategies.

## 2. Related Work

**Video question answering** aims to deduce an answer from a given video in response to a natural language question. There are mainly two types of tasks in this domain: the multiple-choice task *MC-VQA* [24, 26] offers several options, typically up to five, for each question and requires selecting the single correct one, while the open-ended task *OE-VQA* [20, 50, 52, 59] provides a global vocabulary of over 1K possible answers and allows for multiple correct answers to a question. In this work, we mainly focus on OE-VQA because it is more challenging and practical [35, 65]. Existing efforts in OE-VQA are oriented towards two directions. The first emphasizes the ability of *understanding*, where the works aim to build strong video-question joint representations with memory networks [11, 41], attention-based fusion modules [23, 32, 50], or large-scale pretrained backbones [25, 26, 45, 52, 53]. By contrast, the second direction emphasizes *reasoning*, where the works explicitly model the objects and their interactions in videos for QA inference, by leveraging hierarchical structures [10, 40, 42] or graph neural networks [19, 49]. Despite significant progress, the insufficient labeling problem, an essential limitation in OE-VQA benchmarks, is always ignored by these works. In this work, we formally reveal this problem and present promising non-manual solutions.

**Knowledge distillation** (KD) [17] is a famous teacher-student learning paradigm in which a student model is trained by imitating the behavior of a trained teacher model. KD is widely applied in model compression [22, 31, 38, 63], domain generalization [29, 30] and transfer learning [51,

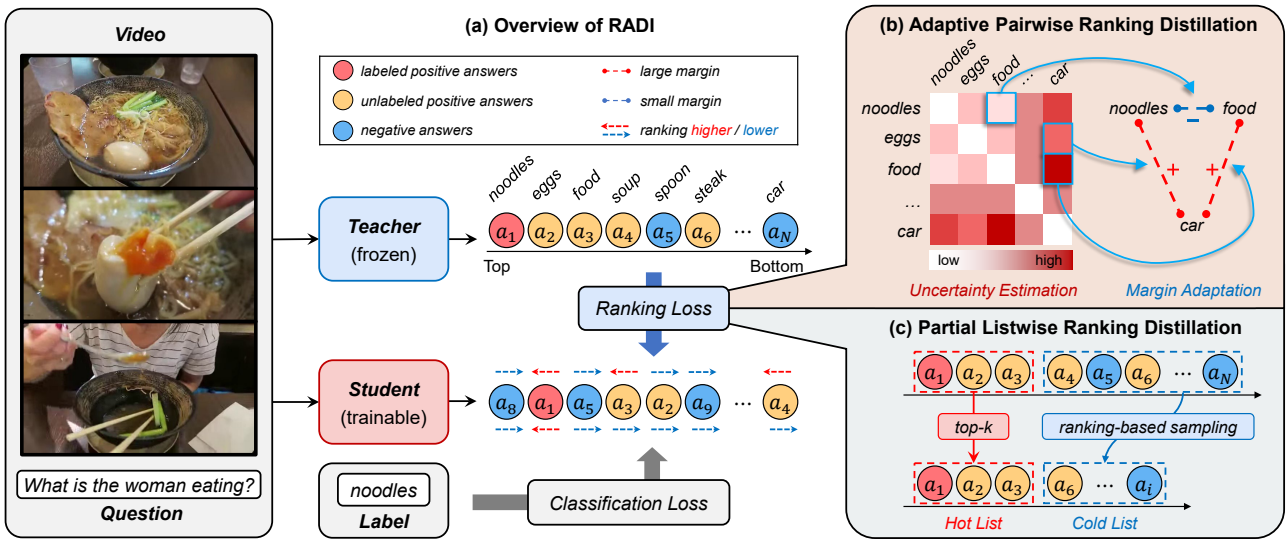


Figure 3. An overview of RADI, which is a LTR-based training framework for OE-VQA. Within RADI, the video-QA model is optimized using two loss functions: (i) **classification loss** maximizes the prediction probability of labeled answer (*i.e.*, *noodles* and *eggs*) and suppress the rest, involving the potential positive answers (e.g., *soup* and *food*); (ii) **ranking loss** may retrieve these potential positive answers, by pushing the predicted ranking to align with the ranking list provided by a well-trained teacher model.

54]. According to the imitating objective, KD has three basic forms [15]: *distribution distillation* [17, 28, 64] distills the output probability distributions, *feature distillation* [39, 47, 60] distills the intermediate features, and *relation distillation* [34, 56, 57] distills the feature relations between different layers or data samples. The last two forms of KD emphasize feature learning, while distribution distillation focus more on label correlation learning. However, all these distillation methods could be suboptimal for the insufficient labeling problem, since they typically treat the teacher model as an infallible oracle, which is hard to achieve particularly in OE-VQA. In this work, we overcome this limitation by introducing a slack ranking distillation framework RADI with two robust learning strategies, which distills the label information in a soft learning-to-rank manner. This manner reduces the sensitivity to the biased prediction and enhances the robustness to noisy knowledge of the imperfect teacher model.

**Learning to rank** is a task aiming to train a model for ranking a list of objects. In general, LTR models can be trained by *pointwise*, *pairwise* or *listwise* methods. Pointwise methods [5, 8, 33] treat each item individually with regression or classification. These method ignore the relationship among different items, typically leading to suboptimal performance. Therefore, recent works in LTR concentrate more on pairwise and listwise learning. Pairwise methods [2, 44, 62] model the relative orders within individual pairs of items, while listwise methods [4, 48] directly model the absolute order of the entire list. In this work, we propose a ranking distillation framework, which formulates the conventional teacher-student learning as a LTR paradigm. Furthermore, we propose two robust LTR approaches to mitigate the impact of the teacher model's bias in distillation.

## 3. Ranking Distillation for OE-VQA

In this work, we propose a ranking distillation framework RADI to overcome the insufficient labeling problem in OE-VQA. As illustrated in Figure 3(a), RADI employs a teacher model, which is merely trained with incomplete labels, to generate informative answer rankings, and then utilize these rankings as external labels to train the student model (Section 3.1). To enhance the robustness of RADI to noisy rankings, we design two alternate distillation methods: *adaptive pairwise ranking distillation*, as shown in Figure 3(b), which adaptively adjusts the pairwise margins based on the teacher model's uncertainties about its pairwise rankings (Section 3.2); and *partial listwise ranking distillation*, as shown in Figure 3(c), which performs partial listwise learning with a novel sampling strategy (Section 3.3).

### 3.1. Overall Framework of RADI

**Task formulation.** Given an answer set $\mathbb{A}$, OE-VQA is formulated as a multi-label classification task, which requires the model to find the correct answers from $\mathbb{A}$ in response to a pair of video $v$ and question $q$. Let $x$ denote a video-question pair, and $p_\theta(a_i|x)$ denote the prediction of a model $\theta$ for each individual answer $a_i$, then the normalized scores are achieved as follows:

$$P_\theta(a_i|x) = \frac{\exp\left(p_\theta\left(a_i|x\right)\right)}{\sum_{j=1}^{N}\exp\left(p_\theta\left(a_j|x\right)\right)}, \tag{1}$$

where $N$ is the size of $\mathbb{A}$.

**Ranking distillation.** The training procedure of RADI consists of two stages. In the first stage, given a sample $x$ and the corresponding labeled set $\mathcal{A}_x$, we train a teacher model $\mathcal{T}$ with the classification loss $\mathcal{L}_{cls}$ as follows:

$$\mathcal{L}_{cls}(\mathcal{T}|\mathcal{A}_x) = -\frac{1}{|\mathcal{A}_x|}\sum_{a\in\mathcal{A}_x}\log P_{\mathcal{T}}(a|x) \tag{2}$$

Once the training of the teacher model is completed, we can obtain a ranked answer list $\mathcal{R}_x$ for each sample $x$, by sorting the teacher's predicted scores over all answers. In the second stage, we utilize both the original labels $\mathcal{A}_x$ and the ranking labels $\mathcal{R}_x$ to jointly train a student model $\mathcal{S}$:

$$\mathcal{L}_x = \mathcal{L}_{cls}(\mathcal{S}|\mathcal{A}_x) + \alpha\mathcal{L}_{rank}(\mathcal{S}|\mathcal{R}_x), \tag{3}$$

where $\mathcal{L}_{rank}$ is the ranking distillation loss that enables the student model to fit the teacher's answer ranking.

Even with only the incomplete labels for training, the teacher model is still able to implicitly learn the label similarities as well as the associations between visual cues and labels owing to the extensive training data. This knowledge depicted in the teacher model can in turn enrich the original limited priori information, contributing to the training of the student model. However, an inevitable challenge arises from the bias and noise inherent in the teacher's knowledge. Directly using existing LTR methods for ranking learning is suboptimal, since they typically require perfect ranking labels. To mitigate this challenge, we further design two robust distillation approaches in the subsequent sections.

### 3.2. Adaptive Pairwise Ranking Distillation

The pairwise ranking approach aims to match the pairwise priorities between the teacher and the student models. This can be accomplished by the *margin ranking loss* as follows:

$$\mathcal{L}_p = \frac{1}{|\mathcal{R}_p|}\sum_{(a_i,a_j)\in\mathcal{R}_p}\max\{0, m-(p_{\mathcal{S}}(a_i)-p_{\mathcal{S}}(a_j))\} \tag{4}$$

where $\mathcal{R}_p = \{(a_i,a_j)|p_{\mathcal{T}}(a_i) > p_{\mathcal{T}}(a_j)\}$ is the set of positive pairwise priorities in the teacher model, $p_{\mathcal{T}}(\cdot)$ and $p_{\mathcal{S}}(\cdot)$ are non-normalized prediction probabilities of the teacher and student model, and $m$ is a constant hard margin.

However, such hard-margin pairwise learning could be susceptible to noisy pairs, since it enforces the student model to match each pairwise priority from the imperfect teacher model without discrimination. To tackle this problem, we introduce adaptive soft margins that dynamically adjust the pairwise constraints based on the teacher model's uncertainty in its predictions. For example, if the teacher model shows high uncertainty on a pairwise priority, then we should relax the learning for this pair, *i.e.*, using a small margin. To this end, we propose soft pairwise ranking distillation, which enhances the robustness to noisy priorities by minimizing the disturbance of uncertain pairs. Specifically, our approach is composed of two steps: the first step aims to estimates the teacher's pairwise uncertainties, and the second step optimizes the pairwise margins for minimizing the overall uncertainty.

**Algorithm 1:** Sinkhorn algorithm for pairwise margin adaptation.

**Input:** Uncertainty matrix $\boldsymbol{U}$, smoothing factor $\lambda$.
**Output:** Margin scaling matrix $\boldsymbol{W}$.

Initialize $\boldsymbol{u}, \boldsymbol{v} \leftarrow \mathbf{1}/N$, $\boldsymbol{v}^{(0)} \leftarrow \mathbf{1}$, $\Delta_v = 1e^8$
$\widetilde{\boldsymbol{U}} \leftarrow \exp(-\boldsymbol{U}/\lambda)$
$\mathrm{diag}(\widetilde{\boldsymbol{U}}) \leftarrow 0$ ; // Ignoring the self-pairs.
**while** $\Delta_v > 1e^{-8}$ **do**
  $\boldsymbol{u}^{(t)} \leftarrow \boldsymbol{u}/(\widetilde{\boldsymbol{U}}\boldsymbol{v}^{(t-1)})$
  $\boldsymbol{v}^{(t)} \leftarrow \boldsymbol{v}/(\widetilde{\boldsymbol{U}}\boldsymbol{u}^{(t)})$
  $\Delta_v \leftarrow \mathrm{mean}(|\boldsymbol{v}^{(t)} - \boldsymbol{v}^{(t-1)}|)$
$\boldsymbol{W} = \mathrm{diag}(\boldsymbol{u}^{(t)})\widetilde{\boldsymbol{U}}\,\mathrm{diag}(\boldsymbol{v}^{(t)})$

**Uncertainty estimation.** Following recent works [7, 55] in uncertainty learning, we employ *Monte Carlo Dropout* [13] to describe the uncertainty of the teacher model. In particular, for each sample, we forward the teacher model with dropout multiple times to derive multiple stochastic predictions. For the sake of efficiency, we only activate the dropout at the last layer, which implies the repetitive execution is only confined to the last layer. Let $p^k(a)$ denote the prediction score for an answer $a$ at $k$-th time, then we can compute the pairwise difference $\boldsymbol{D}^k_{ij} = p^k(a_i) - p^k(a_j)$ at each time. Finally, the uncertainty of the teacher model is represented by the variance of $\boldsymbol{D}$ over $T$ times as follows:

$$\boldsymbol{U} = \frac{1}{T}\sum_{k=1}^{T}(\boldsymbol{D}^k - \bar{\boldsymbol{D}})^2, \tag{5}$$

where $\bar{\boldsymbol{D}} = \frac{1}{T}\sum_{k=1}^{T}\boldsymbol{D}^k$ is the mean matrix. We term $\boldsymbol{U}$ as the *uncertainty matrix*, whose element $\boldsymbol{U}_{ij}$ quantifies the uncertainty of the teacher model on the corresponding pairwise priority $(a_i \to a_j)$.

**Margin adaptation.** The optimal margins of various pairs are determined by minimizing the overall uncertainty, since lower uncertainty tends to indicate higher reliability. This purpose can be achieved by formulating it as an *Optimal Transport* (OT) problem [6, 14, 43]. Specifically, we regard each answer as a sender/receiver, and define the transmission cost from a sender $a_i$ to a receiver $a_j$ as their pairwise uncertainty $\boldsymbol{U}_{ij}$. In this way, our original target can be converted to a standard OT problem—*finding the optimal transport plan with minimal total cost*, which can be efficiently addressed by off-the-shelf OT solvers, such as the Sinkhorn algorithm [9]. We provide the pseudocode in Algorithm 1. Finally, we replace the hard margin $m$ in Equation (4) with the achieved soft margins $\boldsymbol{M}_{ij} = m \cdot \boldsymbol{W}_{ij}$.

### 3.3. Partial Listwise Ranking Distillation

The listwise ranking approach enables the model to directly learn the global ranking. As aforementioned, directly enforcing the student model to strictly match the rankings of the teacher model could be suboptimal, since the teacher model is trained with incomplete labels. To reduce the sensitivity to the biased rankings, we propose to distill only a partial ranking list rather than the entire. However, it is challenging to determine an appropriate ranking sublist for the partial distillation. Indeed, this can be considered from two aspects. On the one hand, the answers near the top of the entire list are significant because they get the highest attention from the teacher model. On the other hand, it is necessary to sample answers broadly from the entire list to support the global ranking. In light of these considerations, we propose a ranking-based sampling strategy, as depicted in Figure 3(c). Firstly, we pick the top-$k$ answers from the teacher model's ranking list as the *hot list*, and sample multiple answers from the remaining list to form the *cold list*. In building the cold list, we employ two alternative ranking-based sampling schemes: *Exp-sampling* formulates the sampling probability of $k$-th answer as: $p_k \propto e^{-\alpha k}$, and *Zipf-sampling* formulates the sampling probability as: $p_k \propto k^{-\alpha}$, where $\alpha$ is a smoothing coefficient. Then, the hot list and the cold list are combined in sequence to derive the desired sublist. Within the sampled subset, we can perform listwise learning to encourage the student model to mimic the teacher model's partial rankings.

Our partial listwise ranking distillation is characterized by high degree of flexibility, enabling adaptation to any standard listwise loss functions, such as:

**ListMLE** [48] directly maximizes the likelihood of the target permutation $\mathcal{R}$ based on the Plackett–Luce model [37]:

$$\begin{aligned}\mathcal{L}_{\text{listmle}} &= -\log P(\mathcal{R}|\mathcal{S}) \\ &= -\log \prod_{i=1}^{n} \frac{\exp\left(p_{\mathcal{S}}(a_{\mathcal{R}_i})\right)}{\sum_{k=i}^{n} \exp\left(p_{\mathcal{S}}(a_{\mathcal{R}_k})\right)},\end{aligned} \tag{6}$$

where $n$ denotes the size of the sublist, and $a_{\mathcal{R}_i}$ denotes the answer ranked at $i$-th position in $\mathcal{R}$.

**ListNet** [4] considers the probability of each answer being ranked at the top (termed *top-1 probability*), and thereby simplifying the lisewise learning into minimizing the cross entropy between the predicted scores and the target scores:

$$\mathcal{L}_{\text{listnet}} = -\sum_{i=1}^{n} \phi\left(p_{\mathcal{T}}(a_i)\right) \log \phi\left(p_{\mathcal{S}}(a_i)\right), \tag{7}$$

where $\phi(\cdot)$ is the *softmax* function.

**STListNet** [1] extend ListNet from deterministic matching to stochastic matching by introducing random disturbance:

$$\mathcal{L}_{\text{stlistnet}} = -\sum_{i=1}^{n} \phi\left(p_{\mathcal{T}}(a_i) + \epsilon_i\right) \log \phi\left(p_{\mathcal{S}}(a_i)\right), \tag{8}$$

where $\epsilon_i = -\beta \log(-\log u_i)$ and $u_i \sim \text{Uniform}(0, 1)$.

**LambdaLoss** [46] is a family of metric-driven listwise loss functions, which can be uniformly defined as:

$$\mathcal{L}_{\text{lambda}} = -\sum_{(a_i, a_j) \in \mathcal{R}_p} \log_2 \left( \frac{1}{1 + e^{-\sigma(p_{\mathcal{S}}(a_i) - p_{\mathcal{S}}(a_j))}} \right)^{w_{ij}}, \tag{9}$$

where $\sigma$ is a hyperparameter, and $w$ is called *lambda weight*. Varying the definition of $w$ leads to various forms of LambdaLoss, such as **RankNet** [3], **NDCG-Loss1**, **NDCG-Loss2** and **NDCG-Loss2++**. Refer to [46] for more details.

## 4. Experiments

In this section, we detail the experimental setup, and then conduct extensive experiments to demonstrate: (1) RADI achieves state-of-the-art performance on multiple mainstream OE-VQA datasets; (2) RADI outperforms other schemes on the insufficient labeling problem in OE-VQA. We also provide ablations to justify the design choices of our method, and present qualitative analyses to show the effectiveness of RADI.

### 4.1. Experimental Setup

**Datasets.** We select five famous OE-VQA datasets for evaluation: **iVQA** [52], ActivityNet-QA [59], **MSVD-QA** [50], **MSRVTT-QA** [50] and **TGIF-FrameQA** [20]. Note that except for iVQA which annotates five ground-truth answers for each sample, all other datasets provide only one answer per sample. Therefore, we use iVQA as the main dataset for evaluation on the insufficient labeling problem.

**Metrics.** In this work, we are concerned about not only the correctness of the answer with the highest predicted score, but also the ranked positions of all correct answers. Hence for a comprehensive evaluation, we apply three distinct metrics: **Acc@1**, which indicates the top-1 accuracy; **Hit@5**, which checks whether at least one ground-truth answer appears within the top-5 prediction; and **nDCG@5** [21], a popular ranking metric that measures the ranking quality at top-5 prediction by considering both the predicted positions and scores of the ground-truth answers. Note that in the comparison experiment with SOTA, we only apply Acc@1 due to the absent results for Hit@5 and nDCG@5 results in prior publications, while for the remaining experiments, we use all the three metrics on iVQA for evaluation.

**Implementation details.** We use FrozenBiLM [53] as the OE-VQA model for both the teacher and student. FrozenBiLM concatenates the frame embeddings and question embeddings in the sequential dimension as the input of a large pretrained language model DeBERTa [16], which then performs visual-textual joint interaction and outputs a matching score for each answer. During training, only a set of lightweight modules like adapters [18] and LayerNorm in FrozenBiLM are updated. For each video, we uniformly sample 10 to 20 frames, with the exact number varying

| Model | #Trainable Params | iVQA | ActivityNet-QA | MSVD-QA | MSRVTT-QA | TGIF-FrameQA |
|---|---|---|---|---|---|---|
| SiaSamRea [58] | - | - | 39.8 | 45.5 | 41.6 | 60.2 |
| Just Ask [52] | 157M | 35.4 | 39.0 | 47.5 | 41.8 | - |
| MERLOT [61] | 223M | - | 41.4 | - | 43.1 | 69.5 |
| VIOLET [12] | 198M | - | - | 47.9 | 43.9 | 68.9 |
| Co-Token [36] | - | 38.2 | - | 48.6 | 45.7 | 62.5 |
| SViTT [27] | 255M | - | 43.2 | - | 43.0 | - |
| All-in-one [45] | 110M | - | - | 48.3 | 46.8 | 66.3 |
| FrozenBiLM [53] | 30M | 39.6 | 43.2 | 54.8 | 47.0 | 68.6 |
| **RADI-P (ours)** | 30M | **43.5** | **44.1** | 55.8 | **48.2** | 69.9 |
| **RADI-L (ours)** | 30M | 42.9 | **44.1** | **56.0** | 48.1 | **70.0** |

Table 1. Comparison with SOTA models on multiple popular OE-VQA datasets. We use **RADI-P** and **RADI-L** to denote the RADI with pairwise ranking distillation and listwise ranking distillation, respectively.

across different datasets. For all datasets, we use AdamW as the optimizer, with a fixed learning rate of $5e^{-5}$ and the linear schedule with warmup. Following [53], we use Dropout with probability of 0.1 in adapters and gradient clipping with the maximum norm of 0.1. In pairwise ranking distillation, we search for the margin scalar $m$ in $\{0.1, 1, 10\}$. In listwise ranking distillation, we normally fix the lengths of the cold list and the hot list as 10 and 100.

### 4.2. Comparison with State-of-the-arts

We compare the top-1 accuracy of our RADI with SOTA models on five popular OE-VQA datasets. As shown in Table 1, both the pairwise version RADI-P and the listwise version RADI-L consistently outperform the existing models on all the datasets. Especially on iVQA that provides multiple ground-truth answers, RADI-P and RADI-L improve the previous SOTA by 3.9% and 3.3% in terms of top-1 accuracy, respectively. This also suggests the effectiveness of our RADI on addressing OE-VQA with insufficient labels. Furthermore, both RADI-P and RADI-L maintain the same number of trainable parameter as FrozenBiLM, owing to our parameter-free pairwise and listwise LTR approaches. In addition, although RADI-P and RADI-L learn to rank from different views, they achieve comparable performance on most of the datasets. This indicates both RADI-P and RADI-L are able to take full advantage of the label information in the ranking list.

### 4.3. Evaluation on Insufficient Labeling Problem

To demonstrate the effectiveness of RADI on the insufficient labeling problem in OE-VQA, we compare it with other potential schemes, as discussed in Section 1, including *label smoothing*, *pseudo labeling* and various *KD* methods on iVQA dataset in terms of Acc@1, Hit@5 and nDCG@5. For KD, we compare with a feature distillation method *FitNet* [39], a relation distillation method *RKD* [34], and three distribution distillation methods: *Vanilla KD* [17] (which directly matches the distributions between teacher and student models), *DKD* [64] (which decouples the distillation loss into target and non-target parts), and *CTKD* [28] (which uses an adaptive temperature). As iVQA provides multiple annotated answers in both training and test sets, to better evaluate the insufficient labeling problem, we retain only the most frequently annotated answer for each training sample. In other words, all schemes **in this experiment** is trained with only one positive answer but evaluated with the complete positive answer set. From the results shown in Table 2, we can observe that:

| Scheme | Acc@1 | Hit@5 | nDCG@5 |
|---|---|---|---|
| FrozenBiLM | 40.5 | 64.5 | 49.6 |
| *Label Smoothing* | | | |
| $\sigma = 0.1$ | 40.2 ▼0.3 | 65.0 ▲0.5 | 49.5 ▼0.1 |
| $\sigma = 0.3$ | 40.2 ▼0.3 | 65.8 ▲1.3 | 50.1 ▲0.5 |
| $\sigma = 0.5$ | 38.7 ▼1.8 | 64.4 ▼0.1 | 48.7 ▼0.9 |
| *Pseudo Labeling* | | | |
| Top-3 | 39.9 ▼0.6 | 66.3 ▲1.8 | 50.3 ▲0.7 |
| Top-5 | 39.0 ▼1.5 | 66.1 ▲1.6 | 49.7 ▲0.1 |
| Top-10 | 37.8 ▼2.7 | 64.9 ▲0.4 | 48.5 ▼1.1 |
| *Knowledge Distillation* | | | |
| Vanilla KD [17] | 41.3 ▲0.8 | 66.7 ▲2.2 | 51.2 ▲1.6 |
| FitNet [39] | 39.2 ▼1.3 | 63.2 ▼1.3 | 48.2 ▼1.4 |
| RKD [34] | 39.2 ▼1.3 | 63.8 ▼0.7 | 48.6 ▼1.0 |
| DKD [64] | 41.4 ▲0.9 | 67.4 ▲2.9 | 51.5 ▲1.9 |
| CTKD [28] | 41.3 ▲0.8 | 66.7 ▲2.2 | 51.0 ▲1.4 |
| **RADI-P (ours)** | **43.2** ▲2.7 | 68.3 ▲3.8 | **52.7** ▲3.1 |
| **RADI-L (ours)** | 42.9 ▲2.4 | **68.6** ▲4.1 | **52.7** ▲3.1 |

Table 2. Evaluation on the insufficient labeling problem with iVQA. In label smoothing, $\sigma$ denotes the smoothing amount. In pseudo labeling, we employ the teacher model's Top-$k$ prediction as the pseudo labels, which are then combined with the original labels to train the student model. Note that the Acc@1 results here might slightly vary from that in Table 1, since all schemes in this experiment can only access one labeled answer during training.

**(i)** Slight label smoothing benefits to Hit@5, while heavy smoothing leads to significant performance degradation. In addition, label smoothing always results in a decline in

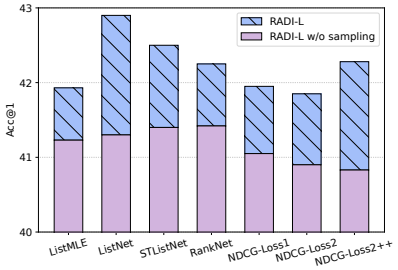


Figure 4. Improvements of using our sampling strategies on various listwise loss functions.

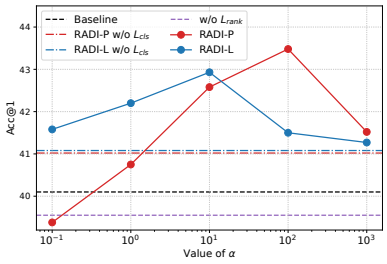


Figure 5. Impacts of using $L_{cls}$, $L_{rank}$ and different $\alpha$ on RADI-P and RADI-L.

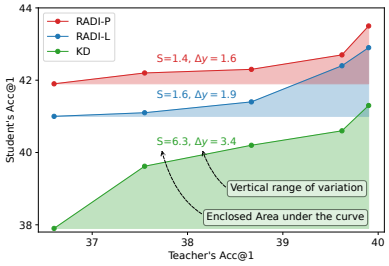


Figure 6. Impacts of using teacher models with different capacities.

Acc@1, since it tends to reduce the confidence of positive answers in exchange for confidence of the negative ones.

**(ii)** Pseudo labeling performs better than label smoothing, but increasing the number pseudo labels leads to a rapid decline in Acc@1, due to the accumulation of noisy labels.

**(iii)** The distribution distillation methods help to allevate the insufficient labeling problem, while the feature distillation method FitNet and the relation distillation method RKD consistently result in performance degradation. This shows the importance of the inter-label information introduced by distribution distillation, as mentioned in Section 2.

**(iv)** Both our RADI-P and RADI-L show great superiority on the insufficient labeling problem, with significant improvements in all the metrics over the baseline schemes. In particular, we find RADI-P performs better in Acc@1 while RADI-L better in Hit@5, and both of them can achieve the best nDCG@5 results. These results demonstrate the effectiveness of our RADI and our improved LTR approaches.

### 4.4. Ablation Study

**Impact of pairwise ranking distillation.** Our adaptive pairwise ranking distillation consists of two steps: uncertainty estimation and OT-based margin adaptation. Hence, we ablate them in this experiment. We design three pairwise baselines: *Uniform* indicates all pairwise priorities share one margin $m$, which is the vanilla pairwise LTR method; *Random* indicates the margin weight of each pairwise priority is a random sampled from $[0, 1]$; and *Uncertainty* indicates directly using the reciprocal of uncertainty as the margin weight. Here we term our pairwise strategy as *Uncertainty+OT*. From the results shown in the upper part of Table 3, we can draw two observations: First, both *Uncertainty* and *Uncertainty+OT* outperform *Uniform* and *Random*, which validates the effectiveness of our first step (*i.e.*, uncertainty estimation). Second, merely using *Uncertainty* would lead to a decline in Acc@1, while *Uncertainty+OT* benefits to all the metrics, which demonstrates the effectiveness of our second step (*i.e.*, OT-based margin adaptation).

**Impact of listwise ranking distillation.** In listwise learning, we propose to distill a partial list rather than the full list, and adopts two sampling strategies to achieve the partial list: *Exp-sampling* and *Zipf-sampling*. In this experiment, we compare them with two baselines: *Full List* indicates distilling the full list, which is the vanilla listwise LTR method, and *Random* indicates that the partial list is obtained by random sampling. All the methods in this experiment adopts ListNet [4] as the loss function. As shown in the lower part of Table 3, the random sampling obtains similar results as *Full List*, while both *Exp-sampling* and *Zipf-sampling* show significant improvements in all the metrics. In addition, we can also see on iVQA dataset, the employed *Zipf-sampling* achieves the best performance.

| Strategy | | Acc@1 | Hit@5 | nDCG@5 |
|---|---|---|---|---|
| Pairwise | *Uniform* | 41.9 | 67.8 | 51.9 |
| | *Random* | 40.9 | 67.8 | 51.6 |
| | *Uncertainty* | 41.5 | 68.2 | 52.2 |
| | *+ OT* | **43.5** | **68.3** | **52.7** |
| Listwise | *Full List* | 41.3 | 66.7 | 51.2 |
| | *Random* | 41.3 | 67.5 | 51.6 |
| | *Exponential* | 42.3 | 68.3 | 52.4 |
| | *Zipf* | **42.9** | **68.6** | **52.7** |

Table 3. Comparison of different strategies of pairwise ranking distillation and listwise ranking distillation on iVQA dataset.

| Dataset | # *answer* | Baseline | +KD | +RADI-L | +RADI-P |
|---|---|---|---|---|---|
| iVQA | 2,349 | 6.5 | 8.0 | 8.0 | 8.2 |
| TGIF-FrameQA | 911 | 8.5 | 11.1 | 11.1 | 11.5 |
| ActivityNet-QA | 1,654 | 18.6 | 23.0 | 23.9 | 25.1 |
| MSVD-QA | 1,198 | 28.5 | 36.3 | 36.3 | 36.6 |
| MSRVTT-QA | 3,589 | 102.0 | 133.2 | 132.9 | 149.6 |

Table 4. Training time (*minutes/epoch*) on a 3090 GPU. For fair comparison, the batch sizes are kept consistent.

**Impacts of various LTR loss functions.** As mentioned in Section 3.3, our partial listwise ranking distillation can be integrated with multiple listwise loss functions. To demonstrate it, we present the results of using various listwise loss functions in Figure 4. It can be observed that our approach achieves consistent improvements across different loss functions, demonstrating the flexibility of our partial listwise ranking distillation.

**Time cost.** Our RADI is an efficient parameter-free training paradigm, which incurs no extra burden at inference and only slightly increases the training time, as shown in Ta-

| Initialization | Acc@1 | Hit@5 | nDCG@5 |
|---|---|---|---|
| *w/o Distillation* | | | |
| From Scratch | 27.6 | 51.0 | 36.9 |
| From Teacher | 40.0 | 66.5 | 50.3 |
| Individual | 40.1 | 65.2 | 49.7 |
| *RADI-P* | | | |
| From Scratch | 42.9 | 68.2 | 52.3 |
| From Teacher | 42.8 | 67.9 | 52.3 |
| Individual | 43.5 | 68.1 | 52.7 |
| *RADI-L* | | | |
| From Scratch | 42.3 | 68.7 | 52.6 |
| From Teacher | 42.4 | 68.3 | 52.5 |
| Individual | 42.9 | 68.6 | 52.7 |

Table 5. Results of different initialization for the student model.

ble 4. The efficiency stems from two aspects. On one hand, RADI-L involves only two additional step—label ranking and sampling—over KD, thus the extra time cost is negligible. On the other hand, RADI-P is efficient for that: 1) Sinkhorn algorithm is an efficient approximation OT solver; 2) MC-dropout is applied to the last layer; 3) the pairwise matrices are truncated up to involving top 2500 answers.

**Impacts of $L_{cls}$ and $L_{rank}$.** In this experiment, we investigate the impacts of $L_{cls}$, $L_{rank}$, and using the different distillation coefficients $\alpha$ selected from $0.1, 1, 10, 100, 1000$. As shown in Figure 5, we can observe that: (i) continual training with only $L_{cls}$ leads to performance decline due to overfitting, while using only $L_{rank}$ can obtain further improvements; (ii) with the increase of $\alpha$, the benefits of RADI-L and RADI-P first increase and then decrease. Specifically, RADI-L performs better than RADI-P when $\alpha$ is small, and achieves the highest Acc@1 at $\alpha = 10$. However, as $\alpha$ continues to rise, RADI-P shows further improvements and achieves the best at $\alpha = 100$.

**Robustness to the imperfect teacher.** We conduct ablation to validate the tolerance of our methods for imperfect teacher models. In this ablation, we evaluate the relation between distillation performance and the capacity of the teacher model, and present the results in Figure 6. As the teacher's performance changes, our RADI-P and RADI-L show more stable performance over KD. Specifically, when Acc@1 of the teacher models decreases by 3.3%, the maximum Acc@1 decline of KD is 3.4%, while that of RADI-P and RADI-L are merely 1.6% and 1.9%, respectively. More significantly, the area under the changing curve of KD is 4.5 and 3.9 times larger than that of RADI-P and RADI-L, respectively. These results demonstrate the benefits of our relaxation strategies to resisting the biased prediction of imperfect teacher models.

**Impact of the student initialization.** There are three ways for initializing our student model: *Scratch*, where the student model is initialized without pretraining on target datasets; *Teacher*, where the student model is initialized with the weights of the trained teacher model; and *Individual*, where the student model is pretrained on the target dataset in the same manner as the teacher model but with a different seed. As shown in Table 5, *Individual* achieves the best performance for both RADI-L and RADI-P, hence we use it within RADI by default. In addition, it is noteworthy that the *Scratch* initialization can still achieve competitive performance, which significantly shows the effectiveness and stability of our approach.

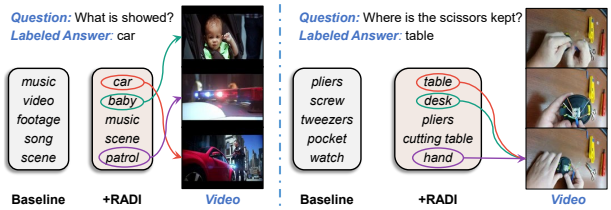


Figure 7. Qualitative results of the top-5 predictions. The arrows in colors connect the correct answers and the visual evidence.

### 4.5. Qualitative Results

We present the qualitative results in Figure 7 by comparing the top-5 predictions between the baseline model FrozenBiLM and our RADI. In the first example, there are at least three correct answers to the question according to the video, while only one answer "car" is labeled. However, RADI not only successfully predicts the labeled answer "car", but also finds two unlabeled correct answers "baby" and "patrol". In the second example, RADI correctly identifys the target position "table", and also finds the synonymous answers "desk" and "cutting table". These two cases show the effectiveness of RADI on the insufficient labeling problem.

## 5. Conclusion

In this paper, we focus on the insufficient labeling problem in OE-VQA task. To alleviate this problem without extra manual annotation, we present a simple but general ranking distillation framework RADI. It employs an imperfect teacher model trained with imcomplete labels to generate answer ranking for enriching the label information. To avoid overconfidence in the imperfect teacher model, we further design two robust and parameter-free ranking distillation approaches: adaptive pairwise ranking distillation with uncertainty-adaptive soft margins, and partial listwise ranking distillation with a ranking-based sampling strategy. We conduct extensive comparison and ablation experiments on five popular OE-VQA datasets to demonstrate the significant improvements of our pairwise and listwise RADIs.

**Acknowledgements**. This work was supported partially by the NSFC (U21A20471, U22A2095, 62076260, 61772570), Guangdong Natural Science Funds Project (2020B1515120085, 2023B1515040025), Guangdong NSF for Distinguished Young Scholar (2022B1515020009), and Guangzhou Science and Technology Plan Project (202201011134).

## References


[1] Sebastian Bruch, Shuguang Han, Michael Bendersky, and Marc Najork. A stochastic treatment of learning to rank scoring functions. In *WSDM*, pages 61–69, 2020. 5

[2] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *ICML*, pages 89–96, 2005. 3

[3] Christopher Burges, Robert Ragno, and Quoc Le. Learning to rank with nonsmooth cost functions. *NeurIPS*, 19, 2006. 5

[4] Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. Learning to rank: from pairwise approach to listwise approach. In *ICML*, pages 129–136, 2007. 3, 5, 7

[5] Rich Caruana, Shumeet Baluja, and Tom Mitchell. Using the future to" sort out" the present: rankprop and multitask learning for medical risk evaluation. In *NeurIPS*, pages 959–965, 1995. 3

[6] Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. Plot: Prompt learning with optimal transport for vision-language models. In *ICLR*, 2022. 4

[7] Daniel Cohen, Bhaskar Mitra, Oleg Lesota, Navid Rekabsaz, and Carsten Eickhoff. Not all relevance scores are equal: Efficient uncertainty and calibration modeling for deep retrieval models. In *ACM SIGIR*, pages 654–664, 2021. 4

[8] Koby Crammer and Yoram Singer. Pranking with ranking. In *NeurIPS*, pages 641–647, 2001. 3

[9] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *NeurIPS*, 26, 2013. 4

[10] Long Hoang Dang, Thao Minh Le, Vuong Le, and Truyen Tran. Hierarchical object-oriented spatio-temporal reasoning for video question answering. In *IJCAI*, pages 636–642, 2021. 2

[11] Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang. Heterogeneous memory enhanced multimodal attention model for video question answering. In *CVPR*, pages 1999–2007, 2019. 2

[12] Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. Violet: End-to-end video-language transformers with masked visual-token modeling. *arXiv preprint arXiv:2111.12681*, 2021. 6

[13] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, pages 1050–1059, 2016. 4

[14] Zheng Ge, Songtao Liu, Zeming Li, Osamu Yoshie, and Jian Sun. Ota: Optimal transport assignment for object detection. In *CVPR*, pages 303–312, 2021. 4

[15] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *IJCV*, 129: 1789–1819, 2021. 3

[16] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*, 2020. 5

[17] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 2, 3, 6

[18] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *ICML*, pages 2790–2799, 2019. 5

[19] Deng Huang, Peihao Chen, Runhao Zeng, Qing Du, Mingkui Tan, and Chuang Gan. Location-aware graph convolutional networks for video question answering. In *AAAI*, pages 11021–11028, 2020. 2

[20] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *CVPR*, pages 2758–2766, 2017. 1, 2, 5

[21] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *TOIS*, 20(4):422–446, 2002. 5

[22] Jangho Kim, SeongUk Park, and Nojun Kwak. Paraphrasing complex network: Network compression via factor transfer. In *NeurIPS*, 2018. 2

[23] Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. Hierarchical conditional relation networks for video question answering. In *CVPR*, pages 9972–9981, 2020. 1, 2

[24] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. Tvqa: Localized, compositional video question answering. In *EMNLP*, page 1369, 2018. 2

[25] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *CVPR*, pages 7331–7341, 2021. 2

[26] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+ language omni-representation pre-training. In *EMNLP*, pages 2046–2065, 2020. 2

[27] Yi Li, Kyle Min, Subarna Tripathi, and Nuno Vasconcelos. Svitt: Temporal learning of sparse video-text transformers. In *CVPR*, pages 18919–18929, 2023. 1, 6

[28] Zheng Li, Xiang Li, Lingfeng Yang, Borui Zhao, Renjie Song, Lei Luo, Jun Li, and Jian Yang. Curriculum temperature for knowledge distillation. In *AAAI*, pages 1504–1512, 2023. 3, 6

[29] Kun-Yu Lin, Jiaming Zhou, Yukun Qiu, and Wei-Shi Zheng. Adversarial partial domain adaptation by cycle inconsistency. In *ECCV*, pages 530–548, 2022. 2

[30] Kun-Yu Lin, Jia-Run Du, Yipeng Gao, Jiaming Zhou, and Wei-Shi Zheng. Diversifying spatial-temporal perception for video domain generalization. In *NeurIPS*, 2023. 2

[31] Sihao Lin, Hongwei Xie, Bing Wang, Kaicheng Yu, Xiaojun Chang, Xiaodan Liang, and Gang Wang. Knowledge distillation via the target-aware transformer. In *CVPR*, pages 10915–10924, 2022. 2

[32] Zihang Lin, Chaolei Tan, Jian-Fang Hu, Zhi Jin, Tiancai Ye, and Wei-Shi Zheng. Collaborative static and dynamic vision-language streams for spatio-temporal video grounding. In *CVPR*, pages 23100–23109, 2023. 2

[33] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Ordinal regression with multiple output cnn for age estimation. In *CVPR*, pages 4920–4928, 2016. 3

[34] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *CVPR*, pages 3967–3976, 2019. 3, 6

[35] Devshree Patel, Ratnam Parikh, and Yesha Shastri. Recent advances in video question answering: A review of datasets and methods. In *ICPR*, pages 339–356, 2021. 1, 2

[36] AJ Piergiovanni, Kairo Morton, Weicheng Kuo, Michael S Ryoo, and Anelia Angelova. Video question answering with iterative video-text co-tokenization. In *ECCV*, pages 76–94, 2022. 6

[37] Robin L Plackett. The analysis of permutations. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 24 (2):193–202, 1975. 5

[38] Antonio Polino, Razvan Pascanu, and Dan Alistarh. Model compression via distillation and quantization. In *ICLR*, 2018. 2

[39] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In *ICLR*, 2014. 3, 6

[40] Chaolei Tan, Zihang Lin, Jian-Fang Hu, Wei-Shi Zheng, and Jianhuang Lai. Hierarchical semantic correspondence networks for video paragraph grounding. In *CVPR*, pages 18973–18982, 2023. 2

[41] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *CVPR*, pages 4631–4640, 2016. 2

[42] Aisha Urooj, Hilde Kuehne, Bo Wu, Kim Chheu, Walid Bousselham, Chuang Gan, Niels Lobo, and Mubarak Shah. Learning situation hyper-graphs for video question answering. In *CVPR*, pages 14879–14889, 2023. 2

[43] Cédric Villani et al. *Optimal transport: old and new*. 2009. 4

[44] Guangcong Wang, Zhaoxi Chen, Chen Change Loy, and Ziwei Liu. Sparsenerf: Distilling depth ranking for few-shot novel view synthesis. In *ICCV*, 2023. 3

[45] Jinpeng Wang, Yixiao Ge, Rui Yan, Yuying Ge, Kevin Qinghong Lin, Satoshi Tsutsui, Xudong Lin, Guanyu Cai, Jianping Wu, Ying Shan, et al. All in one: Exploring unified video-language pre-training. In *CVPR*, pages 6598–6608, 2023. 1, 2, 6

[46] Xuanhui Wang, Cheng Li, Nadav Golbandi, Michael Bendersky, and Marc Najork. The lambdaloss framework for ranking metric optimization. In *CIKM*, pages 1313–1322, 2018. 5

[47] Xionghui Wang, Jian-Fang Hu, Jian-Huang Lai, Jianguo Zhang, and Wei-Shi Zheng. Progressive teacher-student learning for early action prediction. In *CVPR*, pages 3556–3565, 2019. 3

[48] Fen Xia, Tie-Yan Liu, Jue Wang, Wensheng Zhang, and Hang Li. Listwise approach to learning to rank: theory and algorithm. In *ICML*, pages 1192–1199, 2008. 3, 5

[49] Junbin Xiao, Pan Zhou, Tat-Seng Chua, and Shuicheng Yan. Video graph transformer for video question answering. In *ECCV*, pages 39–58, 2022. 1, 2

[50] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *ACM MM*, pages 1645–1653, 2017. 1, 2, 5

[51] Zihui Xue, Sucheng Ren, Zhengqi Gao, and Hang Zhao. Multimodal knowledge expansion. In *ICCV*, pages 854–863, 2021. 2

[52] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *ICCV*, pages 1686–1697, 2021. 1, 2, 5, 6

[53] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Zero-shot video question answering via frozen bidirectional language models. In *NeurIPS*, 2022. 1, 2, 5, 6

[54] Lehan Yang and Kele Xu. Cross modality knowledge distillation for multi-modal aerial view object classification. In *CVPR*, pages 382–387, 2021. 3

[55] Tao Yang, Chen Luo, Hanqing Lu, Parth Gupta, Bing Yin, and Qingyao Ai. Can clicks be both labels and features? unbiased behavior feature collection and uncertainty-aware learning to rank. In *ACM SIGIR*, pages 6–17, 2022. 4

[56] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *CVPR*, pages 4133–4141, 2017. 3

[57] Shan You, Chang Xu, Chao Xu, and Dacheng Tao. Learning from multiple teacher networks. In *ACM SIGKDD*, pages 1285–1294, 2017. 3

[58] Weijiang Yu, Haoteng Zheng, Mengfei Li, Lei Ji, Lijun Wu, Nong Xiao, and Nan Duan. Learning from inside: Self-driven siamese sampling and reasoning for video question answering. In *NeurIPS*, pages 26462–26474, 2021. 6

[59] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *AAAI*, pages 9127–9134, 2019. 1, 2, 5

[60] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *ICLR*, 2016. 3

[61] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. In *NeurIPS*, 2021. 6

[62] Kai Zhang, Chongyang Tao, Tao Shen, Can Xu, Xiubo Geng, Binxing Jiao, and Daxin Jiang. Led: Lexicon-enlightened dense retriever for large-scale retrieval. In *WWW*, page 3203–3213, 2023. 3

[63] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *CVPR*, pages 6848–6856, 2018. 2

[64] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *CVPR*, pages 11953–11962, 2022. 3, 6

[65] Yaoyao Zhong, Wei Ji, Junbin Xiao, Yicong Li, Weihong Deng, and Tat-Seng Chua. Video question answering: Datasets, algorithms and challenges. In *ACL*, pages 6439–6455, 2022. 1, 2