Retrieval-augmented Video Encoding for Instructional Captioning

Yeonjoon Jung[♠] **Minsoo Kim**[♠] **Seungtaek Choi**[♣]
Jihyuk Kim[♡] **Minji Seo**[♠] **Seung-won Hwang**[*♠]
[♠]Seoul National University [♣]Riiid AI Research [♡]Yonsei University
{y970120, minsoo9574, minjiseo, seungwonh}@snu.ac.kr
{seungtaek.choi}@riiid.co {jihyukkim}@yonsei.ac.kr

Abstract

Instructional videos make learning knowledge more efficient, by providing a detailed multimodal context of each procedure in instruction. A unique challenge posed by instructional videos is *key-object degeneracy*, where any single modality fails to sufficiently capture the key objects referred to in the procedure. For machine systems, such degeneracy can disturb the performance of a downstream task such as dense video captioning, leading to the generation of incorrect captions omitting key objects. To repair degeneracy, we propose a retrieval-based framework to augment the model representations in the presence of such key-object degeneracy. We validate the effectiveness and generalizability of our proposed framework over baselines using modalities with key-object degeneracy.

1 Introduction

Instructions, which provide detailed information about the procedures required to achieve the desired goal, are a central part of how humans acquire procedural knowledge. Instructions decompose a sequence of complex procedures into key objects and the associated actions expressed as verbs. As machine systems increasingly aim to provide real-world utility for humans, their ability to translate human goals into natural language instructions to follow becomes essential (Ahn et al., 2022). In this light, instructional captioning, summarizing *instructional videos* into a set of succinct instructions, is thus an important component of enabling the distillation of human-level procedural knowledge to machines.

For instructional captioning, we focus on the task of dense video captioning (DVC) (Krishna et al., 2017) which aims to produce a precise set of instructions from visual input (e.g. instructional videos). For example, to illustrate the procedure s^2 in Figure 1, the instructional video details the procedure, while simultaneously showing how this action is performed. DVC system can then summarize this video into a set of salient captions, forming a set of instructions that enhances the visual demonstration with informative text descriptions.

While the task of extracting a salient instruction from complex visual input can be effortless for humans, it presents a unique challenge for machine systems, which we denote as *key-object degeneracy*. That is, machine systems can often fail at the fundamental task of key-object recognition, which is core to instructions. This is due to the fact that frequently, key objects are not easily recognized from either images (Shi et al., 2019a; Zhou et al., 2018a) or transcripts of the frames (Huang* et al., 2018) during a demonstrative and conversational presentation. While humans can impute such missing information by flexibly aggregating across various available modalities, key-object degeneracy can cause critical failures in existing DVC systems.

Input Modality	Recognizability
Image (X)	56.07
+Transcript (X, T)	63.16
+Instructional Script (X, T, R)	74.60

Table 1: Statistics of the key objects in recognizable forms, recognizability.

To quantify the degeneracy in instructional videos, we first conduct a study measuring the number of recognizable key objects from the images X and transcripts T in one of our target instructional video corpora, YouCook2 (Zhou et al., 2018a)[1]. We define *recognizability* as the percentage of key objects which are recognizable in at least one modality, and present the statistics in Table 1.

From the result in Table 1, we can observe that many key objects are not recognizable from the image alone. Though we can observe that recognizability improves when the image is augmented

*Corresponding author.

[1]We provide detail of computing degeneracy in Sec. 7.2

Findings of the Association for Computational Linguistics: ACL 2023, pages 8554–8568
July 9-14, 2023

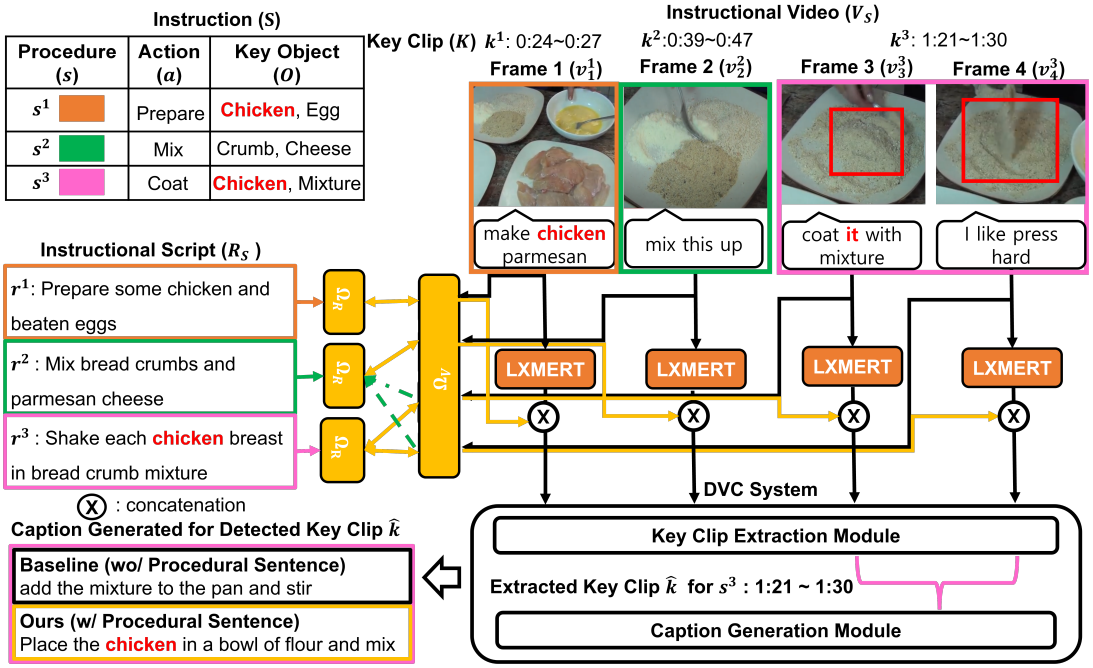

Figure 1: Overall illustration of our framework and a real-life example. The key object "Chicken" of procedure s^3 is hard to recognize from the images and transcripts of Frame 3 and 4 of the instructional video V_S (right top), which we call degeneracy. To repair degeneracy, we supervise the machine system to retrieve procedural sentences (left middle) aligned to each video frame utilizing key object aware inter-frame information (connected with yellow line), unless it fails to distinguish Frame 2 and 3, 4 and retrieve recipe sentence aligned to Frame 2 for Frame 3,4 (connected with green line). We feed frame representation augmented with the retrieved procedural sentence to the downstream task model, DVC, whose generated caption of s^3 (left bottom) becomes more detailed and contains the key object.

with the temporally paired transcript, this does not entirely resolve key-object degeneracy, as nearly 40% of key objects remain unrecognized. For instance, in Figure 1, the key object of procedure s^3, *chicken*, is not recognizable from either the image or transcript of Frame 3.

Having different reasons for degeneracy, each modality has distinct methods to make key objects recognizable: 1) reducing occlusion of key objects in images or 2) reducing ambiguity by mentioning the key objects with nouns in text. Based on the preliminary study, we pursue the latter, and propose a disambiguation method based on retrieval from instructional scripts, such as recipes for cooking.

The sufficient condition of instructional scripts for our method is that they contain *disambiguated* key objects, and provide adequate coverage of valid (key-object, action) pairs. For the YouCook2 dataset, we quantitatively confirm the efficacy of instructional scripts in repairing degeneracy, in Table 1, where it is shown that the instructional scripts can successfully make the unrecognized key objects recognizable. For example, in Figure 1, the unrecognizable key object in the third and fourth frames, *chicken*, becomes recognizable after the procedural sentence $r^3 \in R_S$ (middle left of Figure 1) explicitly mentioning "chicken" is paired with the image and transcript.

While such well-aligned procedural sentences can reduce key-object degeneracy, in most cases, there exists no alignment supervision between the video frame and procedural sentences, as the two are generated independently. Our solution is to generate such alignment using a machine retriever. However, key-object degeneracy in the video frame negatively affects the existing retrieval systems as well, *e.g.*, image-text retrieval, from retrieving the aligned procedural sentence.

Inspired by the contextualized understanding of previous/following frames (Qi et al., 2022), our distinction is to guide the retriever to achieve key-object-aware alignment with procedural sentences, by conducting retrieval based on aggregating inter-frame information in an object-centric manner. For this goal, we propose **K**ey **O**bject aware **F**rame **C**ontrastive **L**earning (KOFCL) for improved differentiation of nearby frames of distinctive procedures, and more robust contextualization of the key object beyond a single procedure.

Our major contributions are threefold: 1) pro-

pose a temporal description retrieval task to find the procedural sentences procedurally aligned to each frame in instructional videos, 2) propose a key object-aware frame contrastive learning objective (KOFCL) to improve temporal description retrieval, and 3) show the improved temporal description retrieval repairs degeneracy and improves DVC significantly.

2 Preliminaries and Related Work

We first introduce our target domain, namely, instruction, and its representations and previous research on their characteristics (§2.1). Our goal is to improve the encoding of video frame G (§2.2). Then we provide a concise overview of our downstream task, DVC (§2.3).

2.1 Target Domain: Instruction and Video, Script

Instruction Instruction refers to structured knowledge explaining how to perform a wide variety of real-world tasks. An instruction S can be represented as a list of N procedures, $S = \{s^j\}_{j=1}^{N}$, where each procedure describes the action required for the task, as a tuple of verb a^j and key object set $\hat{O}^j$, $s^j = (a^j, O^j)$. For example, the instruction for cooking *chicken parmesan* would be a list composed of tuples such as (coat, [chicken, mixture]) which is written in text or shown in the video for human consumption as depicted in Figure 1.

Instructional Video Instructional video, denoted as V_S, is a video explaining instruction S. It consists of a list of frames, $V_S = \{v_i^j | i \leq |V_S| \text{ and } j \leq N\}$. The procedure s^j is represented in the key clip k^j, the subset of video frames starting at b^j and ending at e^j. Then, the i-th frame, v_i^j, represents the corresponding procedure s^j when it is included in the key clip k^j or the null procedure s^0 if it is not covered by any key clip. For example, Frame 1 in Figure 1 explains its procedure by showing and narrating its key objects in its image x_i^j and transcript t_i^j.

It is widely known that degeneracy is prevalent in each modality of instructional videos (Zhou et al., 2018a). Specifically, this indicates a large difference between the key object set O^j and the key objects recognizable in the frame v_i^j, $\hat{O}_i^j$. There have been previous works that discovered and addressed the degeneracy in a single modality of image (Shi et al., 2019b) or transcript (Huang* et al., 2018). However, our approach aims to repair the degeneracy in both modalities, by leveraging the procedural sentences from instructional transcripts.

Instructional Script An instructional script $R_S = \{r^j\}_{j=1}^{N}$ consists of procedural sentences where each procedural sentence r^j represents its corresponding procedure s^j explicitly as words describing the action a^j and the key objects O^j. Representing procedures in disambiguated form, previous works construct instruction S from its corresponding instructional script R_S (Lau et al., 2009; Maeta et al., 2015; Kiddon et al., 2015). We propose to adopt R_S to disambiguate the unrecognizable key object for mitigating degeneracy.

2.2 Baseline: Representation g_i^j

A baseline to overcome degeneracy is to encode the temporally paired image and transcript (x_i^j, t_i^j) into joint multi-modal representation g_i^j. For such purpose, we leverage pretrained LXMERT (Tan and Bansal, 2019)[2], as it is widely adopted to encode the paired image transcript of video frame (Kim et al., 2021; Zhang et al., 2021). Specifically, the transcript t_i^j and image x_i^j of the video frame v_i^j are fed together to pretrained LXMERT. We utilize the representation at the special [CLS] token as the frame representation g_i^j as follows:

$$g_i^j = LXMERT(x_i^j, t_i^j). \quad (1)$$

We use the resulting representation $G = \{g_i^j | i \leq |V_S| \text{ and } j \leq N\}$ as features of individual frames that will be fed to DVC systems.

2.3 Target Task: DVC

Given an instructional video V_S describing instruction S, DVC consists of two subtasks of key clip extraction and caption generation.

Key Clip Extraction Given a sequence of video frames, key clip extraction module predicts key clip $\hat{k} = (\hat{b}, \hat{e})$ by regressing its starting/ending time $\hat{b}$ and $\hat{e}$ (Zhou et al., 2018a; Wang et al., 2021). It also outputs the likelihood $P_k(\hat{k})$ estimating the predicted clip $\hat{k}$ to be a key clip which is further used to select the key clips for caption generation.

Caption Generaton The caption generation task aims to generate caption $\hat{c}$ describing the predicted key clip $\hat{k}$. The predicted key clip $\hat{k}$ is fed to the

[2]We refer to a survey (Du et al., 2022) for overview of multi-modal representation techniques, as our focus is not on enhancing multi-modal representation.

captioning module which generates each word $\hat{w}_i$ by estimating the probability distribution over vocabulary set W conditioned on key clip $\hat{k}$:

$$\hat{w}_i = argmax_{w \in W} P(w|w_{\leq i-1}, \hat{k}). \quad (2)$$

We adopt EMT and PDVC, DVC systems which are widely adopted or SOTA, as our DVC systems. We refer (Zhou et al., 2018b; Wang et al., 2021) for further details, as our focus is not on improving downstream task models, but on repairing the degeneracy of input instructional videos, which is applicable to any underlying models.

3 Our Approach

Building on preliminaries, we now describe our retrieval augmented encoding framework in detail.

First, we explain how instructional scripts can contribute to repairing the degeneracy (§3.1). Our framework combines a cross-modal TDR module (§3.2), which can aggregate the key objects across frames (§3.3), to build robust multi-modal representations which repair key-object degeneracy.

3.1 Representation Augmentation with Procedural Sentence

Our hypothesis to mitigate degeneracy is that a procedural sentence r_i^j in R_S represent a procedure $\tilde{s}_i^j$ similar to the procedure s^j of each frame v_i^j. Explaining a similar procedure, the key object set $\tilde{O}_i^j$ of r_i^j has common key objects sufficient to repair degeneracy. Our first distinction is to augment the individual frame representation g_i^j with the representation d_i^j of such procedural sentence r_i^j. Thus, when procedural sentence r_i^j is provided with video frame v_i^j, more key objects become recognizable,

$$n(O_i^j \cap O^j) \leq n((O_i^j \cup \tilde{O}_i^j) \cap O^j), \quad (3)$$

and the degeneracy in video frames can be reduced.

3.2 Temporal Description Retrieval (TDR)

Cross-modal Retrieval for Aligning Sentences with Frames The preliminary study in Sec. 3.1 establishes the potential of procedural sentences to repair key-object degeneracy. However, it assumes the ideal scenario where the procedure described by the procedural sentence r^j, matches that of the frame v_i^j, which we call *procedural alignment*. However, such procedural alignment between procedural sentences and frames is not available in practice, as data of the two modalities are generated completely independently.

We, therefore, propose a cross-modal retrieval task, Temporal Description Retrieval (TDR), as a solution to *learn* such procedural alignments. We train a frame-sentence retriever, $\phi(v_i^j, R_S)$ to take the query frame v_i^j from video V_S, and the instructional script R_S as input, and predict, for every procedural sentence $r^i \in R_S$, their relevance.The goal of ϕ is to find the procedural sentence $\hat{r}_i$ which best explains the procedure s^j.

Here, it is important to note that the retrieval task itself is also susceptible to key-object degeneracy, making TDR more challenging. In the presence of key-object degeneracy, single-modality (image or text) encodings can exacerbate this problem, due to a potential information imbalance between the two modalities. Therefore, we formulate the cross-modal TDR as retrieving text encodings using a joint image-text query, using the LXMERT joint image-text representation, g_i^j.

Finally, we augment the feature vector g_i^j of the frame with vector representation d_i^j of the retrieved procedural sentence $\hat{r}_i$ as depicted in Figure 1.

Dense Retrieval for Efficiency There can be several options to implement the frame-sentence retriever $\phi(v_i^j, R_S)$. Existing architectures fall into two categories, cross retrievers and dense retrievers (Humeau et al., 2020). These differ in how the interaction between the query frame v_i^j and the procedural sentence r_l is modeled.

As TDR conducts retrieval for each frame in V_S, efficiency should be prioritized, and we mainly consider the dense retrieval architecture. First architecture, the cross retrieval requires the exhaustive computation of $O(|V_S| \times |R_S|)$ as the v_i^j and r_l interact within a single neural network. However, the dense retrieval conducts the retrieval with little computation cost, at $O(|V_S| + |R_S|)$, by reusing the encoding of the v_i^j and r_l.

Specifically, the dense retriever consists of two distinct encoders Ω_V and Ω_R, which encode the query frame v_i^j and the procedural sentence r_l independently. Then, the interaction between v_i^j and r_l is modeled as a simple dot product operation, resulting in retrieval as follows:

$$\hat{r}_i = \text{argmax}_{r_l} \Omega_V(v_i^j) \cdot \Omega_R(r_l). \quad (4)$$

For training, we adopt the contrastive learning objective (Mnih and Kavukcuoglu, 2013), denoted by $\mathcal{L}_{\text{TDR}}$, that guides the retriever to assign larger relevance for the gold procedural sentence r^+ than

that of negative procedural sentences r^-:

$$\mathcal{L}_{\text{TDR}} = -\log \frac{\exp(\Omega_V(v_i^j)\cdot\Omega_R(r^+))}{\exp(\Omega_V(v_i^j)\cdot\Omega_R(r^+))+\sum\exp(\Omega_V(v_i^j)\cdot\Omega_R(r^-))}, \tag{5}$$

We utilize the caption c^j as the gold procedural sentence r^+, as there is no available gold procedural sentence, and this approach is reported to be effective in previous work (Gur et al., 2021). We also utilize in-batch negatives, treating all other gold procedural sentences representing different procedures from the identical instructional video, as negative procedural sentences.

3.3 Key Object-aware Frame Contrastive Learning (KOFCL)

The key aspect separating instructional videos from standard image-text or textual retrieval is the additional temporal dimension. In order to repair key-object degeneracy, it is critical to aggregate inter-frame information across this temporal dimension. To illustrate, consider the key object of frames 3 and 4 in Figure 1, "chicken", which is not recognizable from either the transcript or the images of Frame 3 and 4, but is clearly recognizable in both image v_1^1 and transcript t_1^1 of Frame 1.

We adopt LSTM as a sequence encoder similar to existing video works (Zhou et al., 2018a) and build LXMERT-I^2 which encodes precedent/following frames, $g_{\prec i}^{\leq j}$ and $g_{\succ i}^{\geq j}$, and outputs the resulting query frame encoding $\overleftrightarrow{g}_i^j$ as follows:

$$\overleftrightarrow{g}_i^j = FCN(\overleftrightarrow{LSTM}(g_i^j, g_{\prec i}^{\leq j}, g_{\succ i}^{\geq j})). \tag{6}$$

However, the locality of the frame-level procedure annotations biases such model to simply encode *temporally local* inter-frame information (Wang et al., 2020), not the key objects. Specifically, the procedures are represented as temporally local frames and such local frames of identical procedures can contribute to repair degeneracy. However, as all local frames are not of identical procedures, e.g. boundaries of the key clips, encoding such frames cannot repair degeneracy and rather confuse the models to consider as the preceding/following procedures. For Frame 3 in Figure 1, temporally local inter-frame information of Frame 2 and 3 is redundant with the given frame, adding little new information. Even worse, confusing that Frame 2 and 3 describe the identical procedure, the model misaligns Frame 3 to the procedural sentence r^2 of the different procedure. On the other hand, identifying the key object which appears in Frame 1, and binding this information into the encoding for Frame 3, would successfully repair the key-object degeneracy of Frame 3.

A recent approach, frame contrastive learning (FCL) (Dave et al., 2022), partially addresses the temporal locality bias. It regards the arbitrary frame pair (v_i^j, v_n^m) as positive when they represent identical procedure and negative otherwise as follows:

$$\mathbb{1}(v_i^j, v_n^m) = \begin{cases} 1, & \text{if} j = m \\ 0. & \text{otherwise} \end{cases} \tag{7}$$

What makes FCL address the temporal locality bias is that it supervises the difference in the procedures between the local frames so that local frames of different procedures, such as Frame 2 for given Frame 3 in Figure 1, can be less aggregated.

Then, the frame encoder is supervised to map the frames of identical procedures close together in the representation space, while pushing away those of different procedures by FCL loss, $\mathcal{L}_{aux}(v_i^j, v_n^m)$, defined as follows:

$$y_{in} = \sigma(\overleftrightarrow{g}_i^j \cdot W_{aux} \cdot \overleftrightarrow{g}_n^m) \tag{8}$$

$$\mathcal{L}_{aux}(v_i^j, v_l^k) = BCE(\mathbb{1}(v_i^j, v_n^m), y_{in}), \tag{9}$$

where σ is sigmoid function and W_{aux} is parameter of bilinear layer. Finally, the retriever is optimized to simultaneously minimize $\mathcal{L}_{\text{TDR}}$ and $\mathcal{L}_{\text{aux}}$:

$$\mathcal{L} = \mathcal{L}_{\text{TDR}} + \lambda_{\text{aux}}\mathcal{L}_{\text{aux}}, \tag{10}$$

where λ_{aux} is a hyper-parameter weighing contribution of $\mathcal{L}_{\text{aux}}$ during training.

However, FCL is limited to contextualizing local frames of identical procedure as the inter-frame information. To extend such contextualization beyond the single procedure, we propose key object-aware frame contrastive learning (KOFCL), which encourages contextualizing the frames of different procedures when they share common key objects, based on a globally shared notion of key objects. The clear advantage of such contextualization is that it enables retrieving the correctly aligned procedural sentence, even when key objects are hardly recognizable in the query frame, by leveraging key-object information. For example, the missing key object "chicken" of Frames 3 and 4 in Figure 1 can be found in Frame 1 of procedure s^1, where Frames 1, 3, and 4 will be encouraged to share similar representations through KOFCL. More concretely, we label the frame pair v_i^j and v_n^m as positive when they have common key objects. To measure how

many key objects a frame pair shares, we computed the intersection of union (IoU) between the key object set of frame pair[3] as follows:

$$\text{IoU}_{obj}(v_i^j, v_n^m) = \frac{n(O^j \cap O^m)}{n(O^j \cup O^m)}. \quad (11)$$

Using $\text{IoU}_{obj}(v_i^j, v_n^m)$, we labeled the frame pair, v_i^j and v_m^m, when they share key objects over predefined threshold μ as follows:

$$\mathbb{1}_{obj}(v_i^j, v_n^m) = \begin{cases} 1, & \text{if IoU}_{obj}(v_i^j, v_n^m) > \mu \\ 0. & \text{otherwise} \end{cases} \quad (12)$$

Converting the FCL label in Eq.(7) into our proposed label in Eq.(12), KOFCL supervises to map frame pair, v_i^j and v_n^m, close when they not only describe the identical procedure but also share key objects. Thus, the retriever can build a more robust understanding of the key objects in the query frame v_i^j with key object aware inter-frame information.

4 Experimental Setup

4.1 Dataset

We used two distinct instructional video datasets, YouCook2 (Zhou et al., 2018a), a dataset of instructional cooking videos and IVD (Alayrac et al., 2017), a dataset of instructional videos with 5 distinct goals such as CPR, jump the car. As each video provides its goals, we collected the instructional scripts by querying its goal to the web archive[4] for YouCook2 following previous work (Kiddon et al., 2015) and the Google search engine for IVD dataset. Our instructional script collection contains an average of 15.33 scripts with 10.15 sentences for each goal in YouCook2 and 1 instructional script with an average of 7.4 sentences for each goal in IVD dataset. We used transcripts generated by YouTube ASR engine following previous works (Xu et al., 2020; Shi et al., 2019a, 2020).[5]

4.2 Evaluation Settings

TDR We evaluated TDR in two distinctive settings to utilize both gold captions and our collected instructional scripts. First, we report the recall metric ($R@K$) of the gold captions, where all the captions in the same video are considered candidates for retrieval. Second, we evaluated TDR performance on our collected instructional scripts using $NDCG_{ROUEGE-L}$ metric (Messina et al., 2021a,b). It replaces the relevance annotation between the query frame and procedural sentences with lexical similarity score, ROUGE-L, between gold captions and procedural sentences. We report each metric on top-1/3/5 retrieval result. Especially, for recall metrics, we mainly considered the top-1 retrieval result as our priority is to address key object degeneracy. Specifically, retrieving sentences of different procedures containing the same key objects may result in a slightly lower R@3,5.

DVC For the caption generation of DVC, following convention (Krishna et al., 2017; Zhou et al., 2018b), we report lexical similarity of generated captions with gold captions, using BLEU@4 (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), CIDEr (Vedantam et al., 2015), and Rouge-L (Lin, 2004), abbreviated as B-4, M, C, and R. For the key clip extraction, we report the average recall of the predicted key clips denoted as AR following convention (Escorcia et al., 2016; Zhou et al., 2018b). For every metric, we provide the average and standard deviation of 5 repetitive experiments.

5 Results

We now present our experimental results, aiming to address each of the following research questions:

RQ1: Is our cross-modal retrieval using joint image-text query more effective than standard retrieval approaches for TDR?

RQ2: Does KOFCL address key-object degeneracy in TDR, and help the retriever to build a robust understanding of key objects?

RQ3: Does retrieval-augmentation using procedural sentences improve DVC by repairing key-object degeneracy?

5.1 RQ1: Effectiveness of joint image-text query formulation for TDR

Query Encoder	Input	R@1	R@3	R@5
BM25	t_i^j	35.02	59.34	74.88
BERT	t_i^j	41.45	72.4	86.95
TERAN	x_i^j	39.73	72.39	86.75
NAAF	x_i^j	39.37	72.89	88.17
LXMERT	x_i^j, t_i^j	47.30	78.50	91.14
LXMERT (NAIVE DISAMB.)	x_i^j, τ_i^j	44.75	77.31	90.42
LXMERT-I^2+KOFCL (Ours)	x_i^j, t_i^j	56.83	84.49	94.45

Table 2: Recall (R@1,3,5) for Youcook2 Retrieval with different query frame modality

[3] Human-annotated key object is limited to subset of videos. Therefore, we applied pos-tagging on the groud-truth caption and filtered out the nouns and proper nouns.

[4] www.allrecipes.com

[5] We provide further details of our datasets in Appendix 7.4

Dataset	YouCook2						IVD		
Metric	NDCG			R@K			R@K		
Query Encoder	K=1	K=3	K=5	K=1	K=3	K=5	K=1	K=3	K=5
LXMERT	39.56	41.93	43.50	47.30	78.50	91.14	30.83	62.10	78.54
LXMERT-I^2	41.90	44.21	45.99	55.24	85.86	95.09	40.35	77.77	89.83
+FCL	42.01	44.25	45.82	55.88	85.55	94.89	40.51	74.46	87.15
+KOFCL (OURS)	42.73	44.92	46.50	56.83	84.49	94.45	43.42	76.58	87.86

Table 3: Temporal description retrieval results ablated on inter-frame information

To verify the effectiveness of our joint image-transcript query formulation for TDR, we compare our approach with baselines consisting of existing textual and image-text retrieval systems as follows:

- BM25 (Robertson, 2009) and BERT (Devlin et al., 2019) are widely used approaches in text retrieval. We adopt them as a baseline using the transcript as a query.
- TERAN (Messina et al., 2021a) and NAAF (Zhang et al., 2022) are the state-of-the-art image-text retrievers. We adopt them as baselines using the image x_i^j as a query.

Table 2 shows TDR result of the baselines and our joint image-text query formulation LXMERT for the YouCook2 dataset. We can observe that baselines using single modality queries, i.e. BM25 or TERAN, are insufficient for finding the aligned procedural sentence, with R@1 score lower than 40%. LXMERT shows higher TDR results with large margins over baselines in every metric, confirming the effectiveness of our proposed joint image-transcript query. For comparison, we also include the TDR result of our full model, which further improves significantly over LXMERT.

Additionally, we compare a straightforward method to repair degeneracy, by disambiguating pronouns in transcripts. Following previous work (Huang* et al., 2018), we use a co-reference module (Gardner et al., 2017) to convert transcripts into their disambiguated versions, τ_i^j. Interestingly, we observe a degradation of TDR in every metric. We hypothesize that the co-reference resolution introduces noise from several sources, including the module's inaccuracy itself, but also incorrect pronoun resolution using key objects belonging to other, adjacent procedures.

5.2 RQ2: KOFCL contextualize key objects and improves TDR.

Next, we evaluate the effectiveness of inter-frame information, in conjunction with KOFCL, in improving the performance of TDR. In Table 3, we report the respective results of TDR on the YouCook2 and IVD datasets, with varying inter-frame information supervision approaches.

First, on both datasets, we observe a large improvement of LXMERT-I^2 over LXMERT, reflecting the importance of inter-frame information for TDR. Next, we focus on the effect of jointly supervising LXMERT-I^2 with FCL or KOFCL. When LXMERT-I^2 supervised by FCL, the increase in R@1 is negligible. In contrast, when supervised with our proposed KOFCL, we can observe a meaningful improvement in R@1, on both datasets. These results indicate that KOFCL improves TDR by capturing key-object aware inter-frame information in a generalizable manner.

Query Encoder	R@1
LXMERT-I^2	55.04
+FCL	55.16
+KOFCL (OURS)	56.99

Table 4: Recall@1 score on the isolated set.

In order to further verify that KOFCL contextualizes key objects and repairs key-object degeneracy, we collect an isolated subset of YouCook2, where nearby frames are prone to confuse frame-sentence retrievers with a temporal locality bias. Specifically, we collect the query frames v_i^j whose corresponding procedure s^j has distinct[6] key objects from neighboring procedures s^{j-1} and s^{j+1}.

We report the R@1 score on this isolated set in Table 4. Whereas FCL fails to improve over LXMERT-I^2, R@1 improves meaningfully when the frame-sentence retriever is supervised with our proposed KOFCL. These results indicate that KOFCL contributes to the contextualization of key objects, and alleviates the temporal locality bias.

5.3 RQ3: Retrieved procedural sentences repair degeneracy and improve DVC

Next, we evaluate the impact of repairing degeneracy on improving downstream task of dense video

[6]We considered procedure s^j to have distinct key objects with neighboring procedures when their IoU_{obj} defined in Eq.(12) is lower than 0.05

DVC Model	EMT					PDVC				
Representation	Captioning				KCE	Captioning				KCE
	M	C	R	B-4	AR	M	C	R	B-4	AR
g_i^j	$7.14_{0.20}$	$18.20_{1.09}$	$20.13_{0.52}$	$0.80_{0.09}$	$65.91_{2.95}$	$6.21_{0.42}$	$28.76_{2.61}$	$14.46_{0.75}$	$1.18_{0.16}$	$17.17_{1.09}$
$g_i^j; d_i^j$ w/ τ_i^j	$8.03_{0.25}$	$21.68_{0.61}$	$21.95_{0.80}$	$1.00_{0.08}$	$66.55_{2.99}$	$6.80_{0.44}$	$31.22_{2.10}$	$15.58_{0.94}$	$1.29_{0.14}$	$19.06_{1.27}$
$g_i^j; d_i^j$ w/ LXMERT-I^2 + KOFCL	$\mathbf{8.37}_{0.25}$	$\mathbf{24.37}_{0.67}$	$\mathbf{22.95}_{0.44}$	$\mathbf{1.40}_{0.17}$	$\mathbf{68.93}_{1.72}$	$\mathbf{7.17}_{0.15}$	$\mathbf{33.86}_{0.78}$	$\mathbf{16.55}_{0.45}$	$\mathbf{1.32}_{0.13}$	$\mathbf{20.16}_{0.83}$

Table 5: **B**LEU-**4**, **M**ETEOR, **C**IDEr, **R**ouge-L for captioning, Average Recall (AR) for Key Clip Extraction (KCE).

captioning, which is the main objective of this work. We evaluate our proposed approach, which uses a trained retriever to retrieve procedural sentences from instructional scripts to augment frame representations, with a baseline without any consideration of key-object degeneracy, as well as an advanced baseline, which augments frame representations using the disambiguated version of the transcript τ_i^j, instead of procedural sentences.

We first report the DVC performance on YouCook2 in Table 5. The advanced baseline, which augments the baseline representation g_i^j with d_i^j using τ_i^j, improves performance on both captioning and key clip extraction, showing that DVC can be improved by augmenting frame representations with disambiguated key-object information. Notably, our proposed framework, which augments using procedural sentences retrieved using the LXMERT-I^2 + KOFCL retriever, significantly outperforms both baselines, on all metrics measured, for both tasks. These results indicate that by repairing key-object degeneracy, our retrieved procedural sentences are a better source to augment frame representations for DVC. Moreover, our augmented representations improve results on both EMT and PDVC downstream models, which confirms that our method can be easily applied to improve standard DVC systems, without dramatic modification of the downstream task models.

Representation	Captioning				KCE
	M	C	R	B-4	AR
g_i^j	$7.14_{0.20}$	$18.20_{1.09}$	$20.13_{0.52}$	$0.80_{0.09}$	$65.91_{2.95}$
$g_i^j; d_i^j$ w/ τ_i^j	$8.03_{0.25}$	$21.68_{0.61}$	$21.95_{0.80}$	$1.00_{0.08}$	$66.55_{2.99}$
$g_i^j; d_i^j$ w/ LXMERT	$7.69_{0.21}$	$20.40_{0.69}$	$21.91_{0.49}$	$1.12_{0.15}$	$66.85_{1.08}$
$g_i^j; d_i^j$ w/ LXMERT-I^2	$7.97_{0.33}$	$21.80_{1.21}$	$22.34_{0.50}$	$1.20_{0.15}$	$67.67_{0.25}$
$g_i^j; d_i^j$ w/ LXMERT-I^2 + KOFCL	$\mathbf{8.37}_{0.25}$	$\mathbf{24.37}_{0.67}$	$\mathbf{22.95}_{0.44}$	$\mathbf{1.40}_{0.17}$	$\mathbf{68.93}_{1.72}$

Table 6: **B**LEU-**4**, **M**ETEOR, **C**IDEr, **R**ouge-L for captioning, Average Recall (AR) for Key Clip Extraction (KCE).

Representation	Captioning			KCE
	M	C	R	AR
g_i^j	$9.2_{0.73}$	$61.69_{3.73}$	$14.88_{0.61}$	$36.07_{2.08}$
$g_i^j; d_i^j$ w/ LXMERT-I^2	$16.01_{1.26}$	$102.65_{6.84}$	$24.52_{1.13}$	$27.83_{0.97}$
$g_i^j; d_i^j$ w/ LXMERT-I^2 + KOFCL	$\mathbf{19.76}_{0.85}$	$\mathbf{123.69}_{4.88}$	$\mathbf{29.79}_{0.96}$	$\mathbf{37.97}_{1.61}$

Table 7: Dense video captioning results on IVD dataset. **M**ETEOR, **C**IDEr, **R**ouge-L for captioning, Average Recall (AR) for Key Clip Extraction (KCE).

Next, we conduct an ablation study of the contribution of each of our framework components. In Tables 6 and 7, we report the results of DVC on YouCook2 and IVD respectively, using the EMT model with various frame-sentence retrievers. The results confirm that the improvement in the retrieval outcomes translates to better downstream performance on DVC, with LXMERT-I^2 and KOFCL meaningfully improving DVC performance on both datasets. Also, our proposed retrieval augmentation method showed more improvement in the IVD dataset than YouCook2. The key difference between the Youcook2 and IVD datasets is that the IVD dataset is composed of more distinctive instructions, such as "jump the car", "re-pot the plant" and "make coffee", than YouCook2, which contains only cooking instructions. For such distinctive instructions, knowing the key objects can act as clarifying information about the instruction and thus can help generate more accurate captions.

Representation	Definite	Degenerative
g_i^j	16.38	13.61
$g_i^j; d_i^j$ w/ LXMERT-I^2 + KOFCL	15.33	17.15

Table 8: CIDEr scores results on definite/degenerative sets.

Finally, to verify that the improvement in DVC performance is attributable to the repair key-object degeneracy, we divided the test set into definite and degenerative sets and compared the results of baseline representation g_i^j and our augmented representation $g_i^j; d_i^j$ w/ LXMERT-I^2 + KOFCL. Specifically, the caption c^j is considered degenerative when the video frames corresponding to the ground-truth key clip k^j have lower than 60% recognizability of image and transcript, and definite when the recognizability is higher than 80%. In Table 8, in contrast to representation g_i^j, whose CIDEr score decreases on the degenerative set, our augmented representation $g_i^j; d_i^j$ w/ LXMERT-I^2 + KOFCL increases the score on the degenerative set, showing that our augmented representation using retrieved procedural sentences is effective in re-

solving the key-object degeneracy in instructional videos.

6 Conclusion

We proposed retrieval-augmented encoding, to complement video frames, by repairing degeneracy and considering correlations between steps. Our evaluation results validated that our proposed framework improves existing DVC systems significantly.

Limitations

Our method overcomes degeneracy in instructional videos under the assumption of the existence of textual instructional scripts describing the exact instructions of instructional videos. Thus, our method is applicable to instructional videos having such recipe documents. However, we note that similar documents exist for various types of instructions other than cooking, such as topics in other datasets (Alayrac et al., 2017), *e.g.*, how to jump start a car, or change a tire.

Acknowledgements

This research was supported by MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2023-2020-0-01789) and grants [NO.2021-0-0268, AI Hub, SNU], [No.2022-0-00077, AI Technology Development for Commonsense Extraction, Reasoning, and Inference from Heterogeneous Data], and [NO.2021-0-01343, AI Graduate School] supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation).

References

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil J Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Mengyuan Yan, and Andy Zeng. 2022. Do as i can, not as i say: Grounding language in robotic affordances.

Jean-Baptiste Alayrac, Josef Sivic, Ivan Laptev, and Simon Lacoste-Julien. 2017. Joint discovery of object states and manipulation actions. In *International Conference on Computer Vision (ICCV)*.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Ishan Dave, Rohit Gupta, Mamshad Nayeem Rizve, and Mubarak Shah. 2022. Tclr: Temporal contrastive learning for video representation. *Computer Vision and Image Understanding*, 219:103406.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yifan Du, Zikang Liu, Junyi Li, and Wayne Xin Zhao. 2022. A survey of vision-language pre-trained models.

Victor Escorcia, Fabian Caba Heilbron, Juan Carlos Niebles, and Bernard Ghanem. 2016. Daps: Deep action proposals for action understanding. In *Computer Vision – ECCV 2016*, pages 768–784, Cham. Springer International Publishing.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. 2017. Allennlp: A deep semantic natural language processing platform.

Shir Gur, Natalia Neverova, Chris Stauffer, Ser-Nam Lim, Douwe Kiela, and Austin Reiter. 2021. Cross-modal retrieval augmentation for multi-modal classification.

De-An Huang*, Shyamal Buch*, Lucio Dery, Animesh Garg, Li Fei-Fei, and Juan Carlos Niebles. 2018. Finding "it": Weakly-supervised, reference-aware visual grounding in instructional videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2020. Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Chloé Kiddon, Ganesa Thandavam Ponnuraj, Luke Zettlemoyer, and Yejin Choi. 2015. Mise en place: Unsupervised interpretation of instructional recipes. In *EMNLP*.

Kyungho Kim, Kyungjae Lee, and Seung won Hwang. 2021. Instructional video summarization using attentive knowledge grounding. In *Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track - European Conference, ECML PKDD 2020, Proceedings*, pages 565–569, Germany.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *International Conference on Computer Vision (ICCV)*.

Tessa Lau, Clemens Drews, and Jeffrey Nichols. 2009. Interpreting written how-to instructions. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence*, IJCAI'09, page 1433–1438, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Hirokuni Maeta, Tetsuro Sasada, and Shinsuke Mori. 2015. A framework for procedural text understanding. In *Proceedings of the 14th International Conference on Parsing Technologies*, pages 50–60, Bilbao, Spain. Association for Computational Linguistics.

Nicola Messina, Giuseppe Amato, Andrea Esuli, Fabrizio Falchi, Claudio Gennaro, and Stéphane Marchand-Maillet. 2021a. Fine-grained visual textual alignment for cross-modal retrieval using transformer encoders. *ACM Trans. Multimedia Comput. Commun. Appl.*, 17(4).

Nicola Messina, Fabrizio Falchi, Andrea Esuli, and Giuseppe Amato. 2021b. Transformer reasoning network for image- text matching and retrieval. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 5222–5229.

Andriy Mnih and Koray Kavukcuoglu. 2013. Learning word embeddings efficiently with noise-contrastive estimation. In *Advances in neural information processing systems*, pages 2265–2273.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Jiyang Qi, Yan Gao, Yao Hu, Xinggang Wang, Xiaoyu Liu, Xiang Bai, Serge Belongie, Alan Yuille, Philip Torr, and Song Bai. 2022. Occluded video instance segmentation: A benchmark. *International Journal of Computer Vision*.

S. Robertson. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Zhiqiang Shen, Jianguo Li, Zhou Su, Minjun Li, Yurong Chen, Yu-Gang Jiang, and Xiangyang Xue. 2017. Weakly supervised dense video captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Botian Shi, Lei Ji, Yaobo Liang, Nan Duan, Peng Chen, Zhendong Niu, and Ming Zhou. 2019a. Dense procedure captioning in narrated instructional videos. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6382–6391, Florence, Italy. Association for Computational Linguistics.

Botian Shi, Lei Ji, Zhendong Niu, Nan Duan, Ming Zhou, and Xilin Chen. 2020. *Learning Semantic Concepts and Temporal Alignment for Narrated Video Procedural Captioning*, page 4355–4363. Association for Computing Machinery, New York, NY, USA.

Jing Shi, Jia Xu, Boqing Gong, and Chenliang Xu. 2019b. Not all frames are equal: Weakly-supervised video grounding with contextual similarity and visual clustering losses. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10436–10444.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Teng Wang, Ruimao Zhang, Zhichao Lu, Feng Zheng, Ran Cheng, and Ping Luo. 2021. End-to-end dense video captioning with parallel decoding.

Zhenzhi Wang, Ziteng Gao, Limin Wang, Zhifeng Li, and Gangshan Wu. 2020. Boundary-aware cascade networks for temporal action segmentation. In *ECCV (25)*, volume 12370 of *Lecture Notes in Computer Science*, pages 34–51. Springer.

Frank F. Xu, Lei Ji, Botian Shi, Junyi Du, Graham Neubig, Yonatan Bisk, and Nan Duan. 2020. A benchmark for structured procedural knowledge extraction from cooking videos.

Kun Zhang, Zhendong Mao, Quan Wang, and Yongdong Zhang. 2022. Negative-aware attention framework for image-text matching. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15640–15649.

Yanhao Zhang, Qiang Wang, Pan Pan, Yun Zheng, Cheng Da, Siyang Sun, and Yinghui Xu. 2021. Fashion focus: Multi-modal retrieval system for video commodity localization in e-commerce. *Proceedings*

of the AAAI Conference on Artificial Intelligence, 35(18):16127–16128.

Luowei Zhou, Chenliang Xu, and Jason J Corso. 2018a. Towards automatic learning of procedures from web instructional videos. In *AAAI Conference on Artificial Intelligence*, pages 7590–7598.

Luowei Zhou, Yingbo Zhou, Jason J. Corso, Richard Socher, and Caiming Xiong. 2018b. End-to-end dense video captioning with masked transformer. *CoRR*, abs/1804.00819.

7 Appendix

7.1 Implementation Details

7.1.1 Temporal Description Retrieval

For temporal description retrieval, we followed the convention of (Krishna et al., 2017; Zhou et al., 2018b; Shi et al., 2019a) and obtained the image frames from the video by down-sampling for every 4.5s. The obtained image frames are then fed to pre-trained object detector (Anderson et al., 2018) to yield the sequence of object region features. For image encoder Ω_v and the text encoder Ω_r, we used the image encoder of pretrained LXMERT and BERT-base-uncased (Devlin et al., 2019), respectively. For training temporal description retrieval, we used one video as a batch, resulting in all the sampled frames and recipe sentences in a batch coming from the same video. We adopt an Adam optimizer with a learning rate of 0.0001. We set the weighing contribution λ_{aux} in Eq. 10 to be 0.05 and the threshold μ for KOFCL to be 0.1, based on validation set result.

7.2 Computation of Recognizability

To compute the joint recognizability of the image and transcript, instructional script, we first computed the recognizability in each modality. In the image, we considered the key objects to be recognizable when they are labeled to be inside the image without occlusion in human annotation (Shen et al., 2017). In the textual modality, transcript and instructional script, the key objects are considered to be recognizable when they are lexically referred in transcripts or instructional scripts. Then, we considered the key objects to be recognizable when they are in the union of the recognizable key object set of each modality.

7.3 Ablation on Sequence Encoder

Here, we show the result of TDR with distinct sequence encoders. In Table 9, LSTM showed the highest R@1 score. While we adopted LSTM as our sequence encoder, our KOFCL is orthogonal to any sequence encoder and can be adapted to any existing sequence encoder.

Sequence Encoder	R@1
CNN	50.65
TRANSFORMER	43.69
LSTM (OURS)	55.04

Table 9: Recall@1 score with different sequence encoder.

7.3.1 Dense Video Captioning

EMT For the key clip extraction task, we follow the convention of (Zhou et al., 2018b) to use 16 different kernel sizes for the temporal convolution layer, *i.e.*, from 3 to 123 with the interval step of 8, which can cover the different lengths. We use a transformer encoder and decoder with 768 inner hidden sizes, 8 heads, and 2 layers which we fed context-aware recipe sentences and video frame features after concatenation. We adopt an AdamW optimizer with learning rate of 0.00001 to train the model. The batch size of training is 12 and we use one RTX2080Ti GPU to train our model.

PDVC We use single transformer models with 768 inner hidden sizes, 12 heads, and 2 layers which we fed context-aware recipe sentences and video frame features after concatenation. We adopt an AdamW optimizer with learning rate of 0.00005 to train the model. The batch size of training is 1 and we use one RTX2080Ti GPU to train our model.

7.4 Dataset

We conducted experiments on the two distinct instructional video datasets, YouCook2 (Zhou et al., 2018a), a dataset of instructional cooking videos and IVD dataset (Alayrac et al., 2017), a dataset of instructional videos with 5 distinct topics.

Though YouCook2 originally provides 2000 videos, as some videos are unavailable on YouTube, we collect the currently available videos, obtaining 1,356 videos. For the dataset split, we follow the original split ratio from (Zhou et al., 2018a) to YouCook2: 910 for training, 312 for validation, and 135 for testing for YouCook2. For the IVD dataset, we used 104 for training, 17 for validation, and 32 for testing.

This split is used for both TDR and DVC. Each

video is labeled with starting and ending times of key clips, and their textual descriptions. For transcripts, we use YouTube's ASR engine. We collected the instructional documents from the web archive[7] for YouCook2 following previous work (Kiddon et al., 2015) and top-1 retrieved result from the google search engine for IVD dataset. Our instructional document collection contains an average of 15.33 documents with 10.15 sentences for YouCook2 dataset and 1 instructional document with 20 sentences for IVD dataset.

7.5 Qualitative Results

Here, we provide the generated result of EMT without/with our retrieved recipes in Figure 2. In all examples, there exist the key objects hardly recognizable from the images which EMT fail to mention in the generated caption. However, our retrieved recipes provide the disambiguated reference of such key objects and enable EMT to generate more accurate caption containing them.

[7] www.allrecipes.com

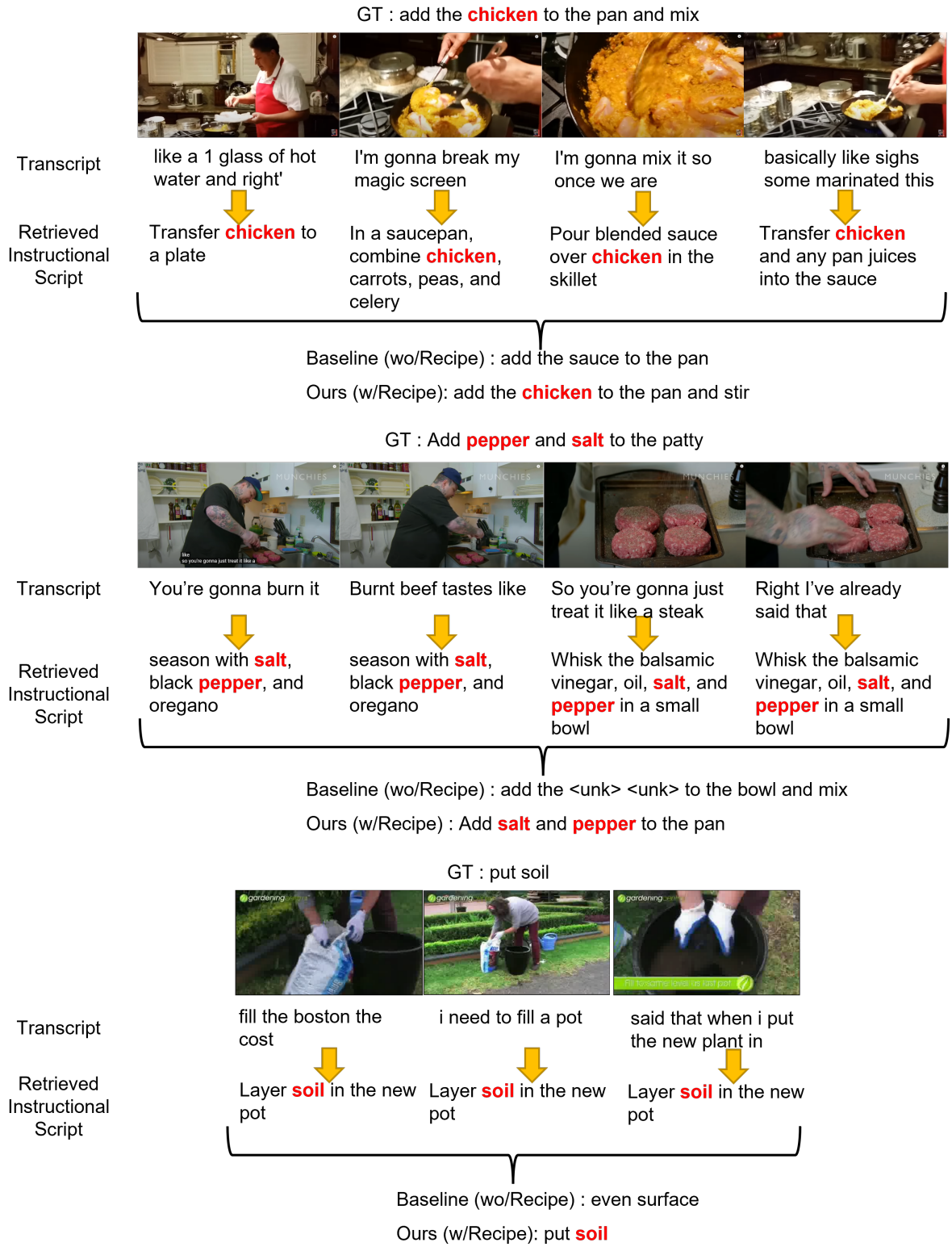

Figure 2: Example of the retrieved procedural sentence and generated captions without/with retrieved procedural sentence. Top 2 figures are from YouCook2 dataset and bottom figure is from IVD dataset.

ACL 2023 Responsible NLP Checklist

A For every submission:

☐ A1. Did you describe the limitations of your work?
Left blank.

☐ A2. Did you discuss any potential risks of your work?
Left blank.

☐ A3. Do the abstract and introduction summarize the paper's main claims?
Left blank.

☐ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☐ Did you use or create scientific artifacts?

Left blank.

☐ B1. Did you cite the creators of artifacts you used?
Left blank.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Left blank.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Left blank.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Left blank.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Left blank.

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Left blank.

C ☐ Did you run computational experiments?

Left blank.

☐ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Left blank.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☐ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Left blank.

☐ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Left blank.

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Left blank.

D ☐ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Left blank.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Left blank.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Left blank.