Procedure Knowledge Decoupled Distillation Strategy for Procedure Planning in Instructional Videos

Xiaotian Pan[1], Zhaobo Qi[1], Xin Sun[1*], Yuanrong Xu[1], Weigang Zhang[1*]

[1]Harbin Institute of Technology, Weihai, China

xiaotianpan@stu.hit.edu.cn, qizb@hit.edu.cn, sunxintyc@hit.edu.cn, xuyuanrong@hit.edu.cn, wgzhang@hit.edu.cn

Abstract

Procedure planning in instructional videos, producing a structured and plannable action sequence facilitating the transition from the start to the goal states, has achieved significant progress. The dominant single-branch non-autoregressive planning paradigm guides action sequence generation through action labels, overlooking the limitation of the absence of intermediate visual information. Hence, we introduce the procedure knowledge decoupled distillation strategy to address the above issue. This innovative strategy deliberately lets the teacher model see the real visual information among the start and goal states to enhance its action semantic understanding and relationship modeling ability, producing the potential probability distribution containing the real action class and other action classes that may occur. Accordingly, we introduce a decoupled intermediate information knowledge distillation loss, which comprises single action knowledge distillation and sequence distribution knowledge distillation for the student model. The former improves the student model's precise inference ability for individual actions by transferring knowledge of a single action target category using binary classification loss. Conversely, the latter uses MSE loss to constrain the student model to learn the action sequence probability distribution from the teacher model, thereby enhancing the student model's global planning capability. Extensive experiments on three datasets demonstrate that our strategy can improve the performance of multiple weakly supervised models, achieving promising procedure knowledge modeling ability and plug-and-play flexibility.

Code — https://github.com/xiaotianpan/PKDD

Introduction

Instructional videos, due to their educational and guiding worth, have garnered considerable attention in the realm of video understanding (Gao, Zhang, and Xu 2020; Tang et al. 2023; Zhong et al. 2023; Gao, Chen, and Xu 2023; Qi et al. 2021, 2024). By analyzing the multimodal content within these videos, humans can easily grasp the sequence of operations to achieve goal-oriented tasks, such as repairing vehicles, cooking, making crafts, etc. While this ability appears intuitive for humans, it poses a formidable challenge for AI systems. Therefore, we concentrate on the procedure planning in instructional video task (Chang et al. 2020). Unlike traditional task planning in structured environments, this endeavor necessitates inferring a structured and plannable action space from unstructured real-world videos that depict the transition from the start state to the goal state given visual observations.

Recently, significant progress has been made in the procedure planning in instructional video task, which can be categorized into two main families: the two-branch autoregressive prediction approach (Chang et al. 2020; Bi, Luo, and Xu 2021; Sun et al. 2022) and the single-branch non-autoregressive prediction method (Zhao et al. 2022b; Wang et al. 2023b; Li et al. 2023). The former considers the auxiliary role of the intermediate information between the start and goal states for action reasoning. It utilizes intermediate action labels and visual states (*i.e.* sampled video frames) for supervised prediction in an iterative inference mechanism (as shown in Figure 1(a)), which tends to result in cumulative errors, with the issue exacerbating in longer sequences. Therefore, the latter adopts a weakly supervised training method to simultaneously predict all intermediate action sequences (as shown in Figure 1(b)) at once.

However, the single-branch non-autoregressive prediction method only incorporates action labels as supervision signals, relying exclusively on the start and goal state information without intermediate visual details. Consequently, it struggles to effectively capture the semantic correlation within the intermediate series of actions, which hinders the model's ability to anticipate actions related to elements absent in the start and target states, limiting the scope of action space to be predicted. For example, as shown in Figure 1, it is difficult for the AI systems to infer the existence of the visual concept "milk" in the intermediate action only from the start and goal observations. Therefore, some weakly supervised models (Wang et al. 2023a; Niu et al. 2024) introduce multiple learning objectives to address this limitation. Nevertheless, they only consider the similarity between individual actions during training, significantly overlooking the intricate spatio-temporal relationship inherent within instructional videos. This oversight manifests in two primary aspects: firstly, the evolution uncertainty of adjacent actions, where even the same ancestral action may have different subsequent ones; secondly, the multifaceted rationality of

*Corresponding author.

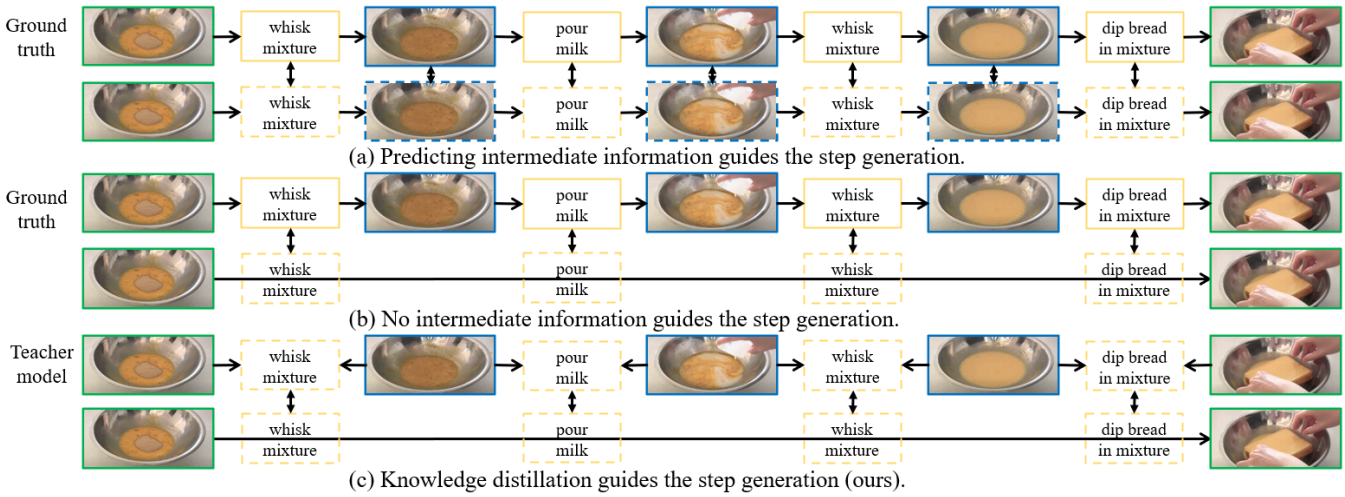

Figure 1: Some research methods for procedure planning. (a) is the step-by-step generation of intermediate states and action sequences by using intermediate information supervision. (b) is the direct generation step using weakly supervised methods. Our method uses a teacher model that can see true intermediate information to supervise step generation, as shown in (c).

intermediate action sequences, where the sequence of intermediate actions does not affect the completion of the final task. Consequently, the model must possess the capacity for long-range global planning while maintaining precise reasoning at the level of individual actions.

To solve the above issues, we propose the procedure knowledge decoupled distillation strategy (PKDD) for procedure planning, as shown in Figure 1(c). Its core idea is to deliberately let the teacher model (based on the diffusion transformer model structure) see the real visual information of the actions among the start and goal observations, which can enhance its action semantic understanding and relationship modeling ability, producing the potential probability distribution of the intermediate action labels. For each intermediate step, compared to the one-hot label, this distribution gives both the probability that the real action class will occur and the probability that other action classes may occur. Considering the evolution uncertainty of the video content, this kind of soft label can assist the student model in learning all potential relationships between actions, thus enhancing its global planning capability and robustness. Based on this, we introduce the Intermediate Information Knowledge Distillation (IIKD), which contains Single Action Knowledge Distillation (SAKD) and Sequence Distribution Knowledge Distillation (SDKD). The former transfers the knowledge of a single action target category to the student model through the binary classification loss to improve the student model's precise inference ability at the individual action level. In contrast, the latter constrains the student model to learn the action sequence probability distribution of the teacher model through the MSE loss, thereby enhancing the student model's global planning ability.

We apply our decoupled distillation strategy to multiple weakly supervised methods and conduct numerous experiments on widely used benchmarks CrossTask (Zhukov et al. 2019), COIN (Tang et al. 2019), and NIV (Alayrac et al. 2016). Experimental results demonstrate that our method achieves more comprehensive procedure knowledge modeling ability and remarkable procedure planning performance. Our main contributions are as follows:

- We propose a procedure knowledge decoupled distillation strategy for procedure planning, which improves the prediction accuracy of the target class and the rationality of the long series distribution of the non-target classes.
- Our strategy can be widely used on multiple weakly supervised methods and has plug-and-play flexibility.
- Extensive experimental evaluation on three widely used datasets demonstrates the effectiveness of our strategy.

Related Work

Procedure Planing

The procedure planning in instructional videos is first introduced in the DNN model (Chang et al. 2020), which uses two-branch autoregression to predict intermediate and action states progressively. Subsequent researches improve upon this by modifying network structures like transformers (Sun et al. 2022) and adversarial policy planning (Bi, Luo, and Xu 2021), resulting in enhanced outcomes. Recent studies model entire action sequences using transformer and diffusion models. Notably, to solve the problem of missing state information, the SCHEMA (Niu et al. 2024) model proposes step representation and state change tracking to handle action planning. The PDPP (Wang et al. 2023b) model uses a diffusion model to generate action sequences directly. However, the lack of intermediate information leads to the decline of the model effect. Therefore, we use the knowledge distillation method to directly utilize the intermediate information and the diffusion model to avoid the complex inference problems caused by multi-objective learning.

Knowledge Distillation

The main idea of the knowledge distillation (Hinton, Vinyals, and Dean 2015) method is to guide the student model to learn from a better and more complex teacher model to improve performance. Some current knowledge distillation methods can be divided into: feature-based (Li et al. 2021) and logit-based (Huang et al. 2022) knowledge distillation. In early research, feature-based methods are better than logit-based methods. Several recent studies (Zhao et al. 2022a; Sun et al. 2024; Wei, Luo, and Luo 2024) have filled these gaps, which leads to a wide range of applications, such as in computer vision (Cui et al. 2023), large language models (Vörös, Bergeron, and Berlin 2023) and other fields (Miles et al. 2023; Patel, Mopuri, and Qiu 2023). Inspired by this, we employ the knowledge distillation technique to extract knowledge from critical action frames within instructional videos. This method aims to reduce the complexity inherent in traditional approaches that rely on intermediate information.

Diffusion Model

In recent years, diffusion probabilistic models have achieved significant success in generative fields. The denoising diffusion probability model (Ho, Jain, and Abbeel 2020) is a generative model based on non-equilibrium thermodynamics, which provides better training stability and higher generation quality. Recent studies emphasize augmenting the sampling speed of diffusion models such as Markov chains (Song, Meng, and Ermon 2020) and optimization of sampling efficiency (Chung, Sim, and Ye 2022). Compared to other generative models, the diffusion model offers notable advantages in representation ability, generalization, and flexibility. These advantages make diffusion models applicable across various domains such as computer vision (Chen et al. 2023; Xia et al. 2023; Kim et al. 2023), natural language processing (Gong et al. 2023a,b), and time series analysis (Rasul et al. 2021). In this paper, we optimize the network structure of the diffusion model based on previous work, which is more helpful for the model to understand the semantic information in the action distribution.

Method

Overview

Task Setting. Given a start observation o_s and a goal observation o_g, the procedure planning in instructional video task aims to forecast a sequence of intermediate actions $\{a_1, ..., a_T\}$ that transforms o_s into o_g, where T is the number of actions (Chang et al. 2020). Following previous research (Wang et al. 2023b), we model the procedure planning problem as an action distribution fitting problem and generate all intermediate actions at once.

Strategy Overview. The overall framework diagram of our procedure knowledge decoupled distillation strategy is shown in Figure 2. Given start and goal observations $\{o_s, o_g\}$, we first obtain the task condition c through the task classifier (a simple MLP model), which is used to guide the generation of intermediate actions.

For the teacher model, which is a diffusion model with a Transformer structure, we first filter out key observation o_k for each intermediate action through semantic similarity comparison. Then, the teacher model produces the probability distribution of intermediate actions $a^R_{1:T}$ with the help of visual observation information about these actions. By letting the teacher model intentionally see these key observations, the teacher model's ability to predict intermediate actions can be enhanced. Finally, we use the ground truth intermediate action labels to train the teacher model through L_{tch}. This process can be expressed as:

$$p(a^R_{1:T}|o_s, o_k, o_g) = \int p(a^R_{1:T}|o_k, c)p(c|o_s, o_g)dc. \quad (1)$$

where $a^R_{1:T}$ is the action sequence from action 1 to action T predicted by the teacher model.

For the student model, which embraces any weakly supervised approach, we refine it utilizing both the original planning loss L_{hard} and our novel intermediate information loss L_{iikd}. This innovative loss comprises the single action knowledge distillation loss and the sequence distribution knowledge distillation loss, which assists the student model in learning the global and local action planning capability from the teacher model for procedure planning. The expression is as follows:

$$p(a^S_{1:T}|o_s, o_g) = \int p(a^S_{1:T}|o_s, o_g, c)p(c|o_s, o_g)dc. \quad (2)$$

where $a^S_{1:T}$ is the action sequence from action 1 to action T predicted by the student model.

Inference Stage. We only use the trained task classifier and student model. The input to the student model comprises the observed data $\{o_s, o_g\}$ alongside the task class c. The outcomes $a^S_{1:T}$ are then adopted as the final results.

Procedure Knowledge Decoupled Distillation Strategy

Teacher Model. Unlike previous research models (Zhao et al. 2022b), we purposely make the real visual observations of all intermediate actions between the start and goal observations available. Hence, the teacher model can extract observational information related to action sequences from large amounts of visual data, which enhances its action semantic understanding and relationship modeling capabilities to accurately infer all intermediate action label distributions.

In detail, since the action clip consists of multiple video frames, some of which may contain redundant information unrelated to the action. We use cosine similarity to compare the similarity of each frame with the semantic representation of the action label. Subsequently, we select a predetermined number of video frames with the highest similarity to represent the key observations o_k of each intermedia action. The key observations o_k and the task class c are concatenated with the noisy action sequence d_N to form the sample x^R_N, where N is the diffusion step and R represents the teacher model. After that, the intermediate action distribution is generated by iterative reverse denoising for x^R_N. Since this paper

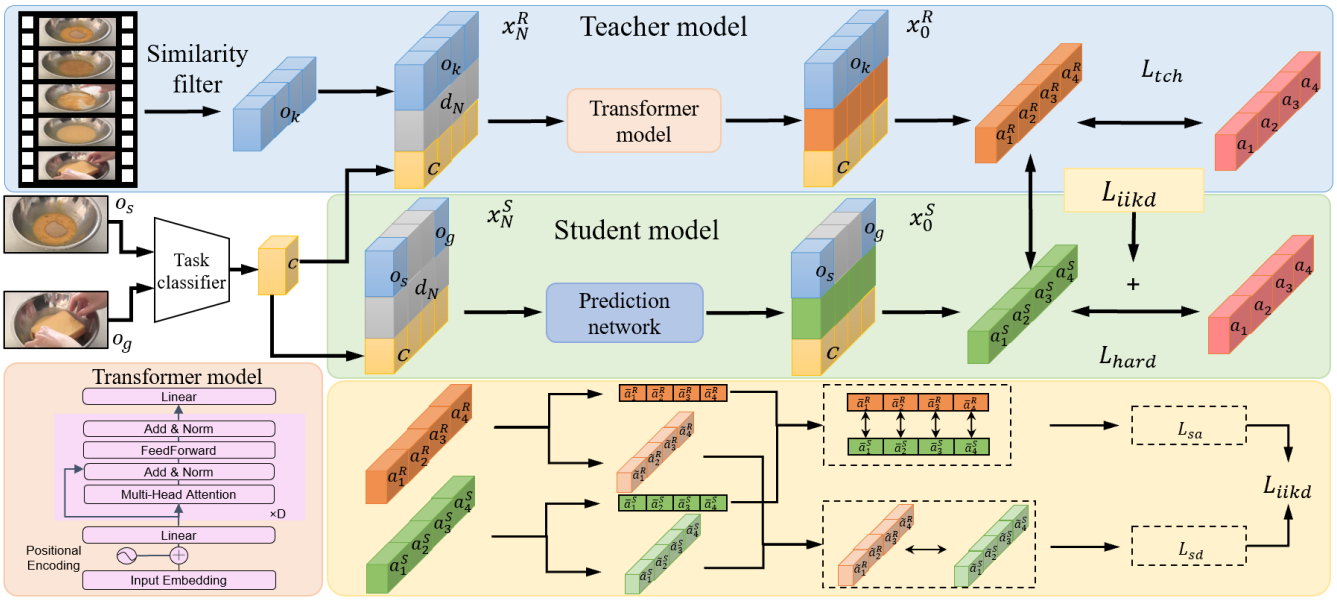

Figure 2: Overview of our PKDD strategy. We let the teacher model see additional key observations o_k to produce the probability distribution of the intermediate action. The student model uses intermediate information knowledge distillation (contains SAKD and SDKD) to learn the potential correlation between actions from this distribution.

needs to fit the distribution to the entire action sequence, directly predicting the noise sampled from random Gaussian distribution will lead to an inability to understand the semantic information (Wang et al. 2023b). Consequently, the learning objective of the teacher model is specified as the sample distribution of the intermediate action x_0^R. In addition, we use the Transformer (Vaswani et al. 2017) structure as the prediction network of the diffusion model, which can better capture the semantic information in the action sequence distribution compared with the U-net (Ronneberger, Fischer, and Brox 2015) structure.

Regarding the constraints imposed on the teacher model, our objective is to let the teacher produce a potential probabilistic distribution of intermediate action labels, which gives the probability that the true action class will occur and the probability that other action classes may occur. Furthermore, we hope that the generated distribution can contain the evolutionary uncertainty of the video content (the probability of all possible subsequent actions of the same ancestor action) to help the student model learn all the potential correlations between actions. Therefore, compared to the original loss, we concatenate the T action distributions predicted by the teacher model to form the entire sequence probability distribution $\{a_{1:T}^R\}$ and align the ground truth intermediate action labels with it to form the sequence label $\{a_{1:T}\}$. We constrain them using the MSE loss, expressed as follows:

$$L_{tch} = \sum_{i=1}^{H} (\{a_{1:T}^R\}_i - \{a_{1:T}\}_i)^2 \tag{3}$$

where H is the number of action sequences.

Student Model. In Figure 2, the PDPP model is used as an example of a student model. The student model concatenates the noised action distribution with $\{o_s, o_g\}$ and c as the noise sample x_N^S, which performs reverse denoising to generate the intermediate action sample x_0^S. For the student model constraints, we utilize the hard labels (the ground truth label of the intermediate actions) and the soft labels (the predicted intermediate actions by the teacher model) to supervise the training of the student model. The strategy only modifies the loss of the student model when using soft label supervision, which can be directly used in various weakly supervised learning methods. Through knowledge distillation, the student model (the original weakly supervised learning model) can greatly enhance its single-action prediction accuracy and process planning capabilities with the help of the teacher model.

Intermediate Information Knowledge Distillation Loss. In knowledge distillation (KD) (Hinton, Vinyals, and Dean 2015), the student model utilizes the hard labels (the ground truth) and the soft labels (predicted by the teacher model) for model training, which distills knowledge from the teacher by minimizing the Kullback-Leibler (KL) divergence between their predictions. This loss is defined as:

$$L_{kd} = KL(\mathbf{p}^R || \mathbf{p}^S)\mathcal{T}^2 \tag{4}$$

where $\mathcal{T}$ is a temperature scaling, $\mathbf{p}^R$ and $\mathbf{p}^S$ are the probability scores output by the teacher and student models.

However, existing knowledge distillation methods cannot adapt to the task, which requires enhancing both the single-action prediction accuracy and the overall sequence planning ability. We hope that the student model can not only accurately predict each action step but also achieve accurate planning of action sequences. Therefore, following (Zhao et al. 2022a), the intermediate information knowledge distillation in our strategy can be decoupled into the weighted

sum of two terms: Single Action Knowledge Distillation (SAKD) and Sequence Distribution Knowledge Distillation (SDKD), where SAKD can focus on the accurate prediction of a single action and SDKD emphasizes the planning ability of the entire action sequence.

In SAKD, we hope that it will focus more on the accuracy of unpredictable actions and help the student model learn unpredictable action information from the teacher model to enhance the student model's accurate inference ability at the individual action level. Therefore, we transfer the single action target class knowledge using the binary cross entropy loss function, which only uses the target class probability predicted by the teacher model to constrain the target class predicted by the student model without paying attention to the probability distribution of the remaining non-target classes. In this way, SAKD can help the student model improve the accuracy of predicting a single action. The loss is as follows:

$$L_{sa} = \sum_{i=1}^{H} (\sum_{j=1}^{T} -(\bar{a}_j^R log(\bar{a}_j^S) + (1 - \bar{a}_j^R) log(1 - \bar{a}_j^S)))_i \tag{5}$$

where $\bar{a}_j^S$ and $\bar{a}_j^R$ are the single target action class probability predicted by the student and teacher model, respectively.

In SDKD, our goal is to enhance the simultaneous fitting of the teacher model's multi-action probability distribution to strengthen the student model's ability to model action relationships and global planning. Like the teacher model, SDKD needs to pay more attention to the accuracy of the entire sequence to ensure that all potential correlations between actions can be learned. Therefore, we concatenate the non-target class distributions of the T actions predicted by the student model to form the entire sequence non-target class probability distribution $\{\tilde{a}_{1:T}^S\}$ and align the teacher model prediction results with it to form the non-target class label $\{\tilde{a}_{1:T}^R\}$. We constrain them using the MSE loss function, which can be integrated into the losses of various weakly supervised methods to enhance the plug-and-play nature of the strategy. The loss is as follows:

$$L_{sd} = \sum_{i=1}^{H} (\{\tilde{a}_{1:T}^S\}_i - \{\tilde{a}_{1:T}^R\}_i)^2 \tag{6}$$

Furthermore, to decouple SAKD and SDKD sufficiently, making the model able to ensure the accuracy of the entire sequence planning while improving the accuracy of single action prediction, we follow previous research (Zhao et al. 2022a) and give different weights to SAKD and SDKD respectively. The soft label loss L_{iikd} is as follows:

$$L_{iikd} = \alpha L_{sa} + \beta L_{sd} \tag{7}$$

We use the original loss function of the student model as the hard label loss, denoted as L_{hard}. Following previous research (Hinton, Vinyals, and Dean 2015), to balance the contribution of IIKD loss and hard label loss when training the student model, we assign certain weights to different losses during training. The final training loss of our student model L_{std} is:

$$L_{std} = \gamma L_{iikd} + \delta L_{hard} \tag{8}$$

where γ and δ are different weights

Experiments

Evaluation Protocol

Dataset. In our evaluation, we employ three datasets originating from distinct sources: CrossTask (Zhukov et al. 2019), COIN (Tang et al. 2019), and NIV (Alayrac et al. 2016). The CrossTask dataset comprises 2,750 video clips categorized into 18 task classes, encompassing a total of 105 action categories. On average, each video clip has 7.6 actions. The COIN dataset is comparatively rich in content, comprising 1,827 video clips, 180 tasks, and 778 action classes. The NIV dataset, although the smallest among the three datasets with 150 videos, 5 tasks, and 18 action classes, boasts the highest action density, averaging 9.5 actions per video clip. We follow previous research to allocate 70% of the data to the training set and the remaining 30% to the test set. All features we use are extracted from the S3D network (Miech et al. 2020) pre-trained on the HowTo100M (Miech et al. 2019) dataset.

Evaluation Metrics. Following previous work (Chang et al. 2020), we use three distinct evaluation metrics. The primary metric employed is the success rate, which is the most rigorous among the metrics used. It deems a model's prediction accurate only if it flawlessly predicts every planned action from the start to the goal state in the correct sequence. The second metric is average accuracy, which evaluates the model's ability to predict a single action correctly. The last metric is the mean Intersection over Union (mIoU), which segregates the predicted actions and actual actions into distinct sets to compute the overlap rate between them. Given that this metric does not account for the specific order of actions, it serves merely as a supplementary measure.

Implementation Details. We adopt traditional settings (Zhao et al. 2022b) to define the start and goal observations for training and inference of the student model in the PKDD strategy. All our experiments are performed on a single GeForce RTX 4090 GPU using the pytorch framework. See the supplementary material for more implementation details.

Comparison with State-of-the-Arts

CrossTask. To verify the effectiveness and plug-and-play nature of our PKDD strategy, we implement it on multiple existing backbones KEPP, SCHEMA, and PDPP. The experiment results on Crosstask datasets are presented in Table 1. Overall, the strategy we proposed can improve the performance of the original weakly supervised model on most indicators of this dataset. This shows that our strategy can effectively utilize intermediate information to improve performance without adding additional inference objectives. From all the experimental results, we find that the performance improvement of the PKDD strategy for different weakly supervised models is different, with the PDPP model showing the highest improvement. According to previous research (Gou et al. 2021), structural differences between teacher and student models affect the effectiveness of knowledge distillation. The structure of the shared diffusion model in PKDD(std$_{PDPP}$) allows the student model to learn

Type	Models	T=3			T=4			T=5	T=6
		SR	mAcc	mIoU	SR	mAcc	mIoU	SR	SR
Autoregressive Model	Random	<0.01	0.94	1.66	<0.01	0.83	1.66	<0.01	<0.01
	Retrieval-Based	8.05	23.30	32.06	3.95	22.22	36.97	2.40	1.10
	WLTDO (Ehsani et al. 2018)	1.87	21.64	31.70	0.77	17.92	26.43	-	-
	UAAA (Abu Farha and Gall 2019)	2.15	20.21	30.87	0.98	19.86	27.09	-	-
	UPN (Srinivas et al. 2018)	2.89	24.39	31.56	1.19	21.59	27.85	-	-
	DDN (Chang et al. 2020)	12.18	31.29	47.48	5.97	27.10	48.46	3.10	1.20
	PlaTe (Sun et al. 2022)	16.00	36.17	65.91	14.00	35.29	55.36	-	-
	Ext-GAIL (Bi, Luo, and Xu 2021)	21.27	49.46	61.70	16.41	43.05	60.93	-	-
Weakly Supervised Learning	KEPP(R=1) (Nagasinghe et al. 2024)	33.34	**61.36**	64.14	20.38	55.54	64.03	13.25	8.09
	KEPP(R=2) (Nagasinghe et al. 2024)	33.38	60.79	63.89	21.02	56.08	64.15	12.74	**9.23**
	PKDD(std$_{KEPP}$)	**33.52**	60.86	**64.30**	**21.22**	**56.42**	**64.31**	**13.38**	8.91
	SCHEMA* (Niu et al. 2024)	31.83	57.31	78.33	19.71	51.85	74.46	11.41	7.68
	PKDD(std$_{SCHEMA}$)	**32.03**	**57.45**	**78.54**	**20.64**	**52.12**	**74.85**	**11.76**	**8.03**
	PDPP (Wang et al. 2023b)	26.38	55.62	59.34	18.69	52.44	62.38	13.22	7.60
	PKDD(std$_{PDPP}$)	**35.47**	**64.54**	**78.19**	**21.61**	**59.20**	**75.57**	**13.74**	**9.59**

Table 1: Evaluation of the results of our model on the CrossTask dataset. * indicates that the SCHAME model additionally introduces language descriptions generated by the large model as input data during the inference process.

Models	T=3			T=4		
	SR	mAcc	mIoU	SR	mAcc	mIoU
SCHEMA*	32.09	49.84	83.83	22.02	45.33	83.47
PKDD(std$_{SCHEMA}$)	**32.77**	**50.47**	**84.32**	**22.79**	**46.33**	**83.91**
KEPP	20.25	39.87	51.72	15.63	39.53	53.27
PKDD(std$_{KEPP}$)	**21.12**	**43.31**	**52.37**	**16.93**	**42.77**	**53.46**
PDPP	20.91	44.81	54.28	14.38	43.63	55.60
PKDD(std$_{PDPP}$)	**27.34**	**55.90**	**63.89**	**20.71**	**54.19**	**66.14**

Table 2: Evaluation results on the COIN dataset.

Models	T=3			T=4		
	SR	mAcc	mIoU	SR	mAcc	mIoU
SCHEMA*	27.93	41.61	**76.77**	**23.26**	**39.93**	**76.75**
PKDD(std$_{SCHEMA}$)	**27.94**	**42.64**	75.97	22.92	38.19	75.62
KEPP	24.44	43.46	86.67	22.71	**41.59**	**91.49**
PKDD(std$_{KEPP}$)	**29.63**	**44.57**	**95.76**	**23.14**	41.38	89.36
PDPP	25.18	42.36	53.67	20.52	41.38	**56.90**
PKDD(std$_{PDPP}$)	**27.04**	**44.81**	**55.40**	**23.14**	**41.92**	55.91

Table 3: Evaluation results on the NIV dataset.

more effectively from the teacher model, while the structural differences in PKDD(std$_{SCHEMA}$) and PKDD(std$_{KEPP}$) reduce this effect.

COIN. Table 2 shows the experimental results of our strategy on the largest COIN dataset. The PKDD strategy improves the performance of different weakly supervised models in all indicators. This shows that our strategy can work better on large-scale datasets and has better generalizability on different datasets.

Std Model Loss	T=3			T=4		
	SR	mAcc	mIoU	SR	mAcc	mIoU
L_{hard}+SAKD	34.96	64.30	**78.29**	20.56	**59.26**	75.51
L_{hard}+SDKD	35.23	64.25	78.02	20.61	59.06	75.08
L_{hard}+IIKD	**35.47**	**64.54**	78.19	**21.61**	59.20	**75.57**

Table 4: Experimental results of student models using different soft label loss functions.

NIV. Table 3 shows the experimental results of our strategy on the NIV dataset. It can be seen that our strategy can improve the performance of the weakly supervised model in most indicators. However, compared with the other two datasets, the performance improvement of the PKDD strategy is not significant and there are certain defects in some indicators. The main reason is that the NIV dataset is too small, which can easily lead to model overfitting problems when performing student model distillation.

Ablation Studies

We conduct ablation experiments on the CrossTask using the PDPP as the student model. Additional ablation experimental study results are provided in the supplementary material.

Different Soft Label Loss Functions. We analyze the impact of different forms of soft label loss on model performance, as shown in Table 4. Among them, L_{hard} refers to the original loss function of the PDPP. From the results, SAKD has a greater impact on the accuracy of the model, while SDKD has a greater impact on the success rate of the model. It can be concluded that SAKD pays more attention to the prediction results of a single action, while SDKD considers overall sequence planning ability. This is consistent with our original design intention.

Methods	T=3			T=4		
	SR	mAcc	mIoU	SR	mAcc	mIoU
Direct Use	33.74	62.61	77.35	19.99	57.11	74.20
KD	35.06	64.09	78.09	20.89	58.75	75.25
KD_{ls}	35.15	64.28	78.17	21.07	58.86	75.49
DKD	35.21	64.22	78.13	21.32	59.08	75.40
DKD_{ls}	35.28	63.95	77.95	21.28	58.93	75.26
IIKD(ours)	**35.47**	**64.54**	**78.19**	**21.61**	**59.20**	**75.57**

Table 5: Experimental results on utilizing intermediate information with different methods. $_{ls}$ means logit normalization is used for preprocessing (Sun et al. 2024).

Methods	T=3			T=4		
	SR	mAcc	mIoU	SR	mAcc	mIoU
No Filter	34.06	63.44	77.71	21.02	**59.45**	75.44
Uniform Sampling	34.98	63.72	77.69	21.27	59.26	75.56
Similarity Filter	**35.47**	**64.54**	**78.19**	**21.61**	59.20	**75.57**

Table 6: Evaluate the impact of the filter on PKDD models.

Methods of Utilizing Intermediate Information. Table 5 shows ablation experiments using different methods to apply intermediate information. Among them, we compare the direct use of the action features extracted by the S3D network (Miech et al. 2020) as intermediate information for supervised training. It can be seen that, compared to the existing knowledge distillation algorithms (such as KD and DKD), the use of IIKD can improve model performance. This proves that intermediate information knowledge distillation can effectively learn the association information between actions to improve the overall planning ability.

Different Filters. Table 6 shows the experimental results of using different filters to filter key observations. Among them, no filter means using the start and goal observations of each action as key observations. From the results, uniform sampling cannot improve model performance very well. The main reason is that action clips containing too much invalid information interfere with the model's extraction of information. The cosine similarity filter can effectively extract the key observations most relevant to the action distribution, reducing the impact of irrelevant observations.

Different Teacher Model Loss. The results in Table 7 show the impact of different loss functions on the performance of the teacher model. The L_{tch} can effectively constrain the teacher model to pay more attention to the connection between actions to improve the relationship modeling capabilities. In addition, the improvement of mAcc proves that L_{tch} can ensure the teacher model in single-action prediction accuracy while paying more attention to the overall prediction effect of multiple actions in a sequence.

Tch Model Loss	T=3			T=4		
	SR	mAcc	mIoU	SR	mAcc	mIoU
Original Loss	59.45	81.39	**88.68**	48.88	79.71	**88.21**
L_{tch}	**60.09**	**81.65**	88.48	**49.68**	**79.65**	87.36

Table 7: Experimental results using different loss functions for the teacher model.

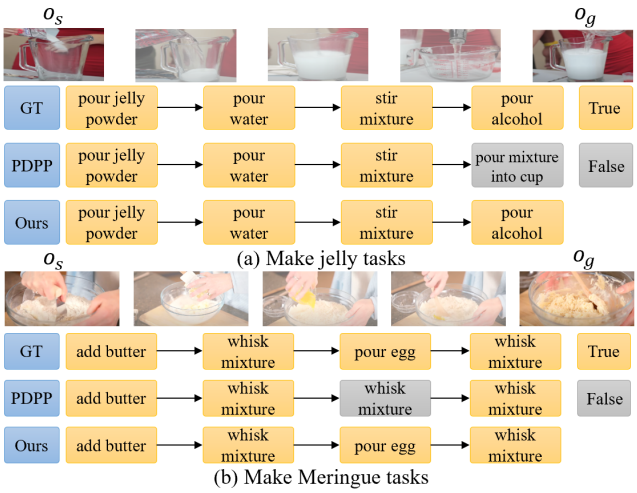

Figure 3: Qualitative analysis in different task situations.

Qualitative Analysis

We take the PDPP as the baseline and conduct a qualitative analysis of our PKDD in different tasks, as shown in Fig 3. Obviously, the PDPP cannot capture intermediate information that does not appear from the given start and goal observations, such as wine and eggs that appear in the intermediate state. This makes the PDPP unable to predict actions that require the use of wine and eggs. The PKDD strategy using the knowledge distillation method can learn intermediate information from the teacher model, so it can successfully predict action steps with unknown information in the middle. This proves that the addition of intermediate information helps the model expand the predicted action space.

Conclusion

In this work, we propose the use of knowledge distillation in instructional videos for procedure planning, aiming to utilize real intermediate information to improve model performance. We call this approach PKDD (Procedure Knowledge Decoupled Distillation). The method lets the teacher model see additional real visual information by filtering key observations and applies this information to train the student model by using a disentangled intermediate information knowledge distillation loss. Furthermore, our strategy has the plug-and-play flexibility to improve the performance of multiple weakly supervised learning models. In future work, we will continue to explore procedure planning with multiple-task fusion in teaching videos to build a more general procedure knowledge distillation strategy.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant U21B2038, 62306092, 62306091, 62072141, and 62476068, and in part by the Natural Science Foundation of Shandong Province, China: ZR2024QF066 and ZR2023QF052.

References

Abu Farha, Y.; and Gall, J. 2019. Uncertainty-Aware Anticipation of Activities. In *2019 IEEE/CVF International Conference on Computer Vision Workshop*, 1197–1204.

Alayrac, J.-B.; Bojanowski, P.; Agrawal, N.; Sivic, J.; Laptev, I.; and Lacoste-Julien, S. 2016. Unsupervised learning from narrated instruction videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4575–4583.

Bi, J.; Luo, J.; and Xu, C. 2021. Procedure planning in instructional videos via contextual modeling and model-based policy learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15611–15620.

Chang, C.-Y.; Huang, D.-A.; Xu, D.; Adeli, E.; Fei-Fei, L.; and Niebles, J. C. 2020. Procedure planning in instructional videos. In *European Conference on Computer Vision*, 334–350.

Chen, S.; Sun, P.; Song, Y.; and Luo, P. 2023. Diffusiondet: Diffusion model for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 19830–19843.

Chung, H.; Sim, B.; and Ye, J. C. 2022. Come-closer-diffuse-faster: Accelerating conditional diffusion models for inverse problems through stochastic contraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12413–12422.

Cui, K.; Yu, Y.; Zhan, F.; Liao, S.; Lu, S.; and Xing, E. P. 2023. Kd-dlgan: Data limited image generation via knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3872–3882.

Ehsani, K.; Bagherinezhad, H.; Redmon, J.; Mottaghi, R.; and Farhadi, A. 2018. Who let the dogs out? modeling dog behavior from visual data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4051–4060.

Gao, J.; Chen, M.; and Xu, C. 2023. Vectorized Evidential Learning for Weakly-supervised Temporal Action Localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12): 15949 – 15963.

Gao, J.; Zhang, T.; and Xu, C. 2020. Learning to model relationships for zero-shot video classification. *IEEE transactions on pattern analysis and machine intelligence*, 43(10): 3476–3491.

Gong, S.; Li, M.; Feng, J.; Wu, Z.; and Kong, L. 2023a. DiffuSeq: Sequence to Sequence Text Generation with Diffusion Models. In *International Conference on Learning Representations, ICLR*.

Gong, S.; Li, M.; Feng, J.; Wu, Z.; and Kong, L. 2023b. DiffuSeq-v2: Bridging Discrete and Continuous Text Spaces for Accelerated Seq2Seq Diffusion Models. *arXiv preprint arXiv:2310.05793*.

Gou, J.; Yu, B.; Maybank, S. J.; and Tao, D. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6): 1789–1819.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Huang, T.; You, S.; Wang, F.; Qian, C.; and Xu, C. 2022. Knowledge distillation from a stronger teacher. *Advances in Neural Information Processing Systems*, 35: 33716–33727.

Kim, M.; Liu, F.; Jain, A.; and Liu, X. 2023. Dcface: Synthetic face generation with dual condition diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12715–12725.

Li, Z.; Geng, W.; Li, M.; Chen, L.; Tang, Y.; Lu, J.; and Zhou, J. 2023. Skip-Plan: Procedure planning in instructional videos via condensed action space learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10297–10306.

Li, Z.; Ye, J.; Song, M.; Huang, Y.; and Pan, Z. 2021. Online knowledge distillation for efficient pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 11740–11750.

Miech, A.; Alayrac, J.-B.; Smaira, L.; Laptev, I.; Sivic, J.; and Zisserman, A. 2020. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9879–9889.

Miech, A.; Zhukov, D.; Alayrac, J.-B.; Tapaswi, M.; Laptev, I.; and Sivic, J. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2630–2640.

Miles, R.; Yucel, M. K.; Manganelli, B.; and Saà-Garriga, A. 2023. Mobilevos: Real-time video object segmentation contrastive learning meets knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10480–10490.

Nagasinghe, K. R. Y.; Zhou, H.; Gunawardhana, M.; Min, M. R.; Harari, D.; and Khan, M. H. 2024. Why Not Use Your Textbook? Knowledge-Enhanced Procedure Planning of Instructional Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18816–18826.

Niu, Y.; Guo, W.; Chen, L.; Lin, X.; and Chang, S.-F. 2024. SCHEMA: State Changes Matter for Procedure Planning in Instructional Videos. *arXiv preprint arXiv:2403.01599*.

Patel, G.; Mopuri, K. R.; and Qiu, Q. 2023. Learning to retain while acquiring: combating distribution-shift in adversarial data-free knowledge distillation. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 7786–7794.

Qi, Z.; Wang, S.; Su, C.; Su, L.; Huang, Q.; and Tian, Q. 2021. Self-regulated learning for egocentric video activity anticipation. *IEEE transactions on pattern analysis and machine intelligence*, 45(6): 6715–6730.

Qi, Z.; Wang, S.; Zhang, W.; and Huang, Q. 2024. Uncertainty-Boosted Robust Video Activity Anticipation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12): 7775–7792.

Rasul, K.; Seward, C.; Schuster, I.; and Vollgraf, R. 2021. Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting. In *International Conference on Machine Learning*, 8857–8868.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, 234–241.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Srinivas, A.; Jabri, A.; Abbeel, P.; Levine, S.; and Finn, C. 2018. Universal planning networks: Learning generalizable representations for visuomotor control. In *International conference on machine learning*, 4732–4741.

Sun, J.; Huang, D.-A.; Lu, B.; Liu, Y.-H.; Zhou, B.; and Garg, A. 2022. Plate: Visually-grounded planning with transformers in procedural tasks. *IEEE Robotics and Automation Letters*, 7(2): 4924–4930.

Sun, S.; Ren, W.; Li, J.; Wang, R.; and Cao, X. 2024. Logit standardization in knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15731–15740.

Tang, Y.; Ding, D.; Rao, Y.; Zheng, Y.; Zhang, D.; Zhao, L.; Lu, J.; and Zhou, J. 2019. Coin: A large-scale dataset for comprehensive instructional video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1207–1216.

Tang, Y.; Liu, J.; Liu, A.; Yang, B.; Dai, W.; Rao, Y.; Lu, J.; Zhou, J.; and Li, X. 2023. Flag3d: A 3d fitness activity dataset with language instruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22106–22117.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Vörös, T.; Bergeron, S. P.; and Berlin, K. 2023. Web content filtering through knowledge distillation of large language models. In *2023 IEEE International Conference on Web Intelligence and Intelligent Agent Technology*, 357–361.

Wang, A.-L.; Lin, K.-Y.; Du, J.-R.; Meng, J.; and Zheng, W.-S. 2023a. Event-Guided Procedure Planning from Instructional Videos with Text Supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13565–13575.

Wang, H.; Wu, Y.; Guo, S.; and Wang, L. 2023b. Pdpp: Projected diffusion for procedure planning in instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14836–14845.

Wei, S.; Luo, C.; and Luo, Y. 2024. Scaled Decoupled Distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15975–15983.

Xia, B.; Zhang, Y.; Wang, S.; Wang, Y.; Wu, X.; Tian, Y.; Yang, W.; and Van Gool, L. 2023. Diffir: Efficient diffusion model for image restoration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13095–13105.

Zhao, B.; Cui, Q.; Song, R.; Qiu, Y.; and Liang, J. 2022a. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 11953–11962.

Zhao, H.; Hadji, I.; Dvornik, N.; Derpanis, K. G.; Wildes, R. P.; and Jepson, A. D. 2022b. P3iv: Probabilistic procedure planning from instructional videos with weak supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2938–2948.

Zhong, Y.; Yu, L.; Bai, Y.; Li, S.; Yan, X.; and Li, Y. 2023. Learning procedure-aware video representation from instructional videos and their narrations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14825–14835.

Zhukov, D.; Alayrac, J.-B.; Cinbis, R. G.; Fouhey, D.; Laptev, I.; and Sivic, J. 2019. Cross-task weakly supervised learning from instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3537–3545.