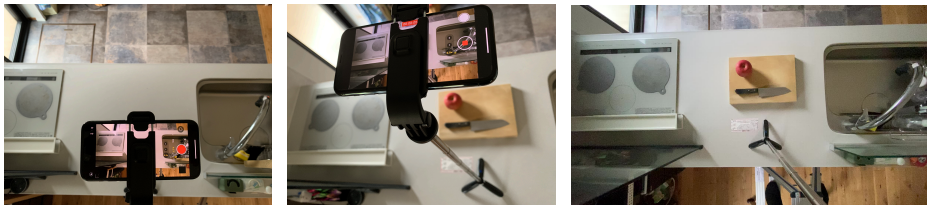
A Detailed Data Appendices

To further aid in understanding, a few examples from COM Kitchens are provided with the supplementary material in the `examples` directory. Samples include (i) unedited recorded videos, (ii) annotations for Japanese recipes, (iii) annotations for translated English recipes, and (iv) constructed visual action graphs. Besides, we also provide a video wall (`videowall.mp4`) to overview the unedited videos, which demonstrates the diversity of our dataset.

B Film set

We provide an example of the film set in Fig. 8. In the recording, we employed a tripod with 900 mm of height and instructed to place it with prior confirmation that the wide-angle mode of the rear camera could cover the whole kitchen top.



(a) Top-down view. **(b)** Angled top-down view. **(c)** The footage of a capture.

Fig. 8: Example of the film set and recorded content.

C Reason of rejection

Table 5: Breakdown of reasons for refusal with statistics.

Reason	# of Refusals	% among Refusals
1. Inappropriate view (e.g., stove is not covered)	97	50.5%
2. Faces in the view	46	24.0%
3. Skipped steps using pre-processed food	12	6.3%
4. Overly complicated process	11	5.7%
5. Pause and resume in recording	9	3.1%
6. Recording by slow mode	5	2.6%
7. Personal documents in the view	3	1.6%
8. Any other reasons	10	3.1%

We summarize the reason for refusal with its statistics in Tab. 5. We had to refuse roughly 50% of the submitted videos (192/412), which is a relatively high rate. Our instructional videos and documents are almost for items 1 and 2, but the ignorance of those instructions caused 74.5% of refusals. This was caused primarily due to the lack of pre-filtering. Since we selected to collect videos with the same smartphone model this time, we had to distribute our equipment to participants, which made pre-filtering difficult.

We judged a procedure too complicated if the video duration was more than one hour or had more than 30 APs or 10 actions in an AP. They were caused by our failure in the recipe selection. In addition, we refused some videos if an actor repeated tasting and adjusting the taste too many times or repeated actions of wrapping small ingredients that were almost invisible in the video.

The other reasons were incomplete information in the consent form (3 videos), withdrawal of consent at the request (2 videos), and removal of the recipe from the Cookpad website (1 video).

D Additional Results on the OnRR task

Tab. 6 lists the result of baseline models in the OnRR benchmark in the early- and middle-stage setting. Tab. 7 showcases the rest results, late-, and full-stage setting. These results suggest that in the recipe stage retrieval task, fine-tuning with our COM Kitchens dataset improves the performance, regardless of model types and cooking stage settings. On the other hand, the reduced and unstable performance in the feasible recipe retrieval task implies that the conventional contrastive learning strategy does not fit the objectives.

Table 6: Online recipe retrieval (OnRR) performances of baseline models **without fine-tuning on COM Kitchens** in the early- and middle-stage settings (using the first 25% and 50% of the video as input). R@K and MdR represent recall at rank K (↑) and median rank (↓), respectively. The results with fine-tuning are shown in Tab. 3.

Task	Method	**Early (25%)**				**Middle (50%)**			
		R@1	R@5	R@10	MdR	R@1	R@5	R@10	MdR
Feasible Recipe Retrieval	Random	1.8	8.6	15.8	-	0.4	1.8	3.1	-
	UniVL [25]	3.4	10.3	17.2	56.0	3.4	10.3	17.2	56.0
	CLIP4Clip [26]	3.4	6.8	13.7	60.0	3.4	3.4	10.3	94.0
	X-CLIP [27]	3.4	10.3	13.7	111.0	0.0	3.4	3.4	569.0
Recipe Stage Identification	Random	6.3	31.6	63.3	8.0	6.3	31.6	63.3	8.0
	UniVL [25]	6.8	37.9	65.5	7.0	0.0	41.3	86.2	5.0
	CLIP4Clip [26]	6.8	31.0	51.7	9.0	3.4	41.3	82.7	7.0
	X-CLIP [27]	6.8	37.9	51.7	8.0	6.8	34.4	51.7	8.0

Table 7: Online recipe retrieval (OnRR) performances of baseline models in late- (75%) and full-stage (100%) settings. The rows with 'FT' of '✓' show the results of models fine-tuned on the COM Kitchens dataset. Note that as the cooking stage progresses, random results in feasible recipe retrieval deteriorate due to the reduced number of feasible recipes.

Task	Method	FT	**Late (75%)**				**Full (100%)**			
			R@1	R@5	R@10	MdR	R@1	R@5	R@10	MdR
Feasible Recipe Retrieval	Random	-	0.0	0.0	0.0	-	0.0	0.0	0.0	-
	UniVL [25]		3.4	10.3	17.2	56.0	3.4	10.3	17.2	56.0
	UniVL [25]	✓	3.4	5.7	9.2	231.0	3.4	5.7	9.2	231.0
	CLIP4Clip [26]		3.4	3.4	10.3	85.0	3.4	3.4	6.8	77.0
	CLIP4Clip [26]	✓	0.0	0.0	6.8	91.0	0.0	0.0	3.4	72.0
	X-CLIP [27]		0.0	0.0	0.0	860.0	0.0	0.0	0.0	911.0
	X-CLIP [27]	✓	0.0	0.0	0.0	446.0	0.0	0.0	0.0	366.0
Recipe Stage Identification	Random	-	6.3	31.6	63.3	8.0	6.3	31.6	63.3	8.0
	UniVL [25]		0.0	41.3	86.2	5.0	0.0	48.2	96.5	5.0
	UniVL [25]	✓	6.8	44.8	86.2	5.0	6.8	51.7	96.5	4.0
	CLIP4Clip [26]		3.4	20.6	79.3	7.0	0.0	24.1	93.1	7.0
	CLIP4Clip [26]	✓	6.8	48.2	93.1	5.0	10.3	55.1	89.6	4.0
	X-CLIP [27]		10.3	41.3	93.1	6.0	6.8	58.6	89.6	4.0
	X-CLIP [27]	✓	10.3	41.3	93.1	6.0	10.3	62.0	89.6	3.0

E Additional Visual Examples on DVC-OV tasks

The following examples are included to provide further insights and reinforce the points made in the main text. Here, we present some more cases in Fig. 9. As with the other cases, we confirm that the combination of supervision connected related frames, using action graphs as relation labels (RL) and as attention supervision (AS).

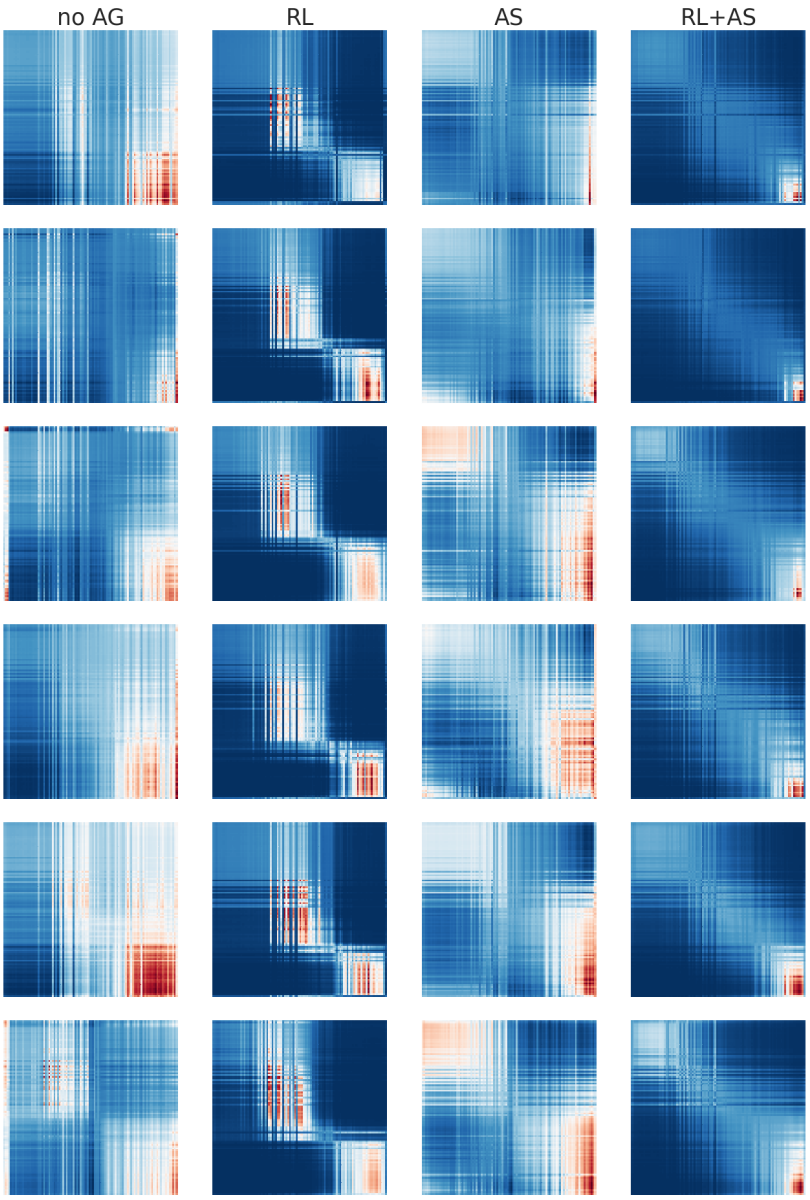


Fig. 9: Additional examples of attention of the first head at the last encoder layer. The red area indicates the high attention weights.