Supplementary - Zero-Shot Anticipation for Instructional Activities

Fadime Sener
University of Bonn, Germany
sener@cs.uni-bonn.de

Angela Yao
National University of Singapore
ayao@comp.nus.edu.sg

1. Tasty Videos Dataset

The videos in the Tasty Videos Dataset are captured with a fixed overhead camera and are focused entirely on the preparation of the dish. Videos are designed to be primarily visually informative without any narrations, except for the textual recipe steps. An example video for "Weekday Meal-prep Pesto Chicken and Veggies" can be found online here: https://tasty.co/recipe/weekday-meal-prep-pesto-chicken-veggies, which is shown in Fig. 24. More videos can be found on https://tasty.co/.

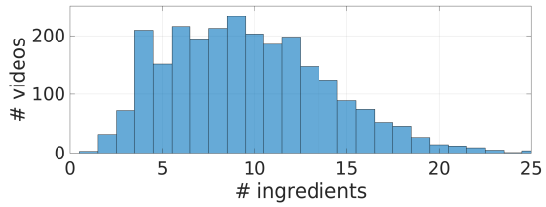

Figure 1: Distribution of the number of ingredients (out of 1199 unique ingredients).

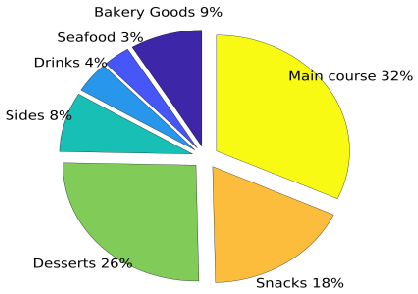

Figure 2: Rough categorization of the recipes in our dataset.

In our dataset, there are 1199 unique ingredients and the average number of ingredients is 9, see Figure 1. In comparison, the number of unique ingredients in the Recipe1M dataset [5] is around 4K. Our dataset has a large variety of meals, including main courses, snacks, sides etc., see Fig. 2.

Each recipe has a list of instructions. For each recipe step, we annotate the temporal boundaries in which the step occurs within the video, omitting those without visual correspondences, such as alternative recommendations. As such, there are less visual instructions than text-based ones.

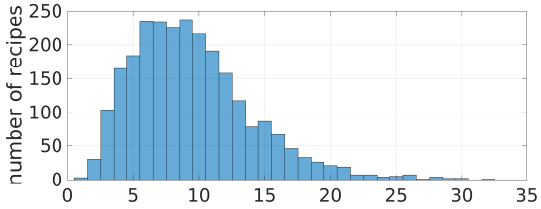

Figure 3: Distribution of the number of visual steps. The average number is 9.

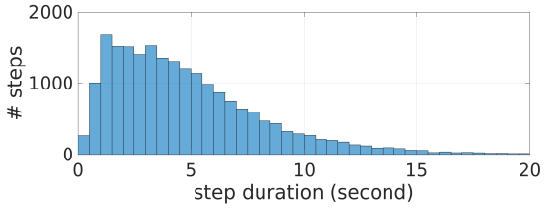

Figure 4: Distribution of annotated visual step durations.

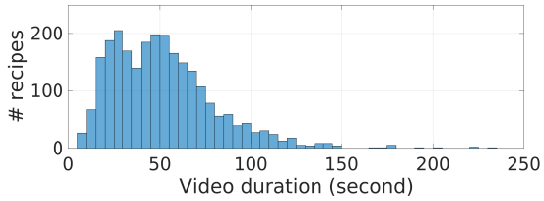

Figure 5: Distribution of video durations.

The average number of visual recipe steps is 9, and there are 21236 visual recipe steps in total. In Figure 3, we show the distribution of the number of visual recipe steps. In Figure 4, we report the distribution of the duration of the annotated visual steps. The average visual step duration is 144 frames or 5 seconds. In Figure 5, we report the distribution of the duration of our videos. The shortest video lasts 6 seconds while the longest lasts 233 seconds. The average video duration is 1551 frames or 54 seconds.

2. Experiments

2.1. Learning of Procedural Knowledge

In our experiments, we first target important keywords, specifically ingredients and verbs, since they indicate the

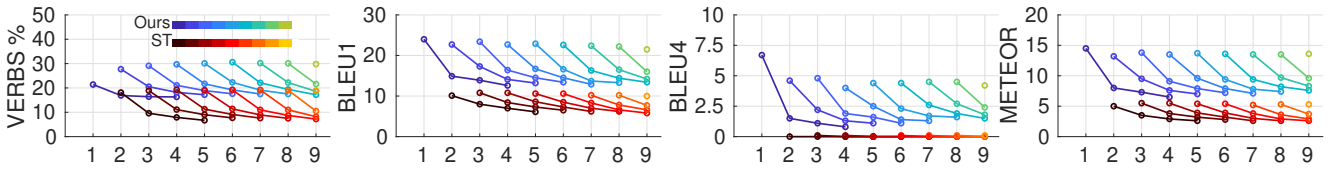

Figure 6: The recall of verbs and sentence scores computed between the predicted and ground truth sentences for our model (Ours) and the skip-thoughts (ST) model over the entire test set of the Recipe1M dataset [5]. The x-axes indicate the step number being predicted in the recipe; each curve begins on the first (relative) prediction, i.e. *the $(j+1)$th step after having received steps 1 to j as input.*

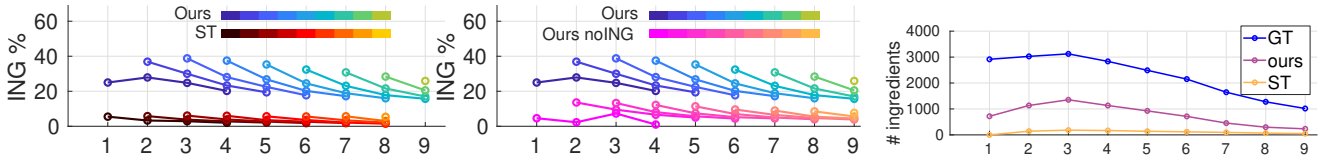

Figure 7: a) The recall of ingredients predicted by our model (Ours), by our model trained without the ingredients (Ours noING) and by skip-thoughts model (ST) over the entire test set of the Recipe1M dataset [5]. The x-axes in the plots indicate the step number being predicted in the recipe; each curve begins on the first (relative) prediction, i.e. *the $(j+1)$th step after having received steps 1 to j as input. b) Absolute number of ingredients detected in the ground truth steps (GT), steps predicted by our model (ours) and the skip-thoughts model (ST) computed over recipes with exactly 9 steps. The number of ingredients detected in a recipe decreases towards the end of the recipe.*

next active object and next actions. Key ingredients and verbs alone do not capture the rich instructional nature of recipe steps, compare *e.g. 'whisk'* and *'egg'* to *'Whisk the eggs till light and fluffy'*. As such, we also evaluate the predicted sentences as a whole and compare to ST predictions based on standard sentence evaluation metrics, such as BLEU (BiLingual Evaluation Understudy) [4] and the METEOR score (Metric for Evaluation of Translation with Explicit ORdering) [1]. BLEU computes an n-gram based precision for predicted sentences with respect to ground truth sentences. METEOR creates an alignment between the ground truth and predicted sentence using the exact word matches, stems, synonyms, and paraphrases; it then computes a weighted F-score with an alignment fragmentation penalty. For the uninformed reader, we note that these scores are best at indicating precise word matches to ground truth. Yet in natural spoken language, much variation may exist between sentences conveying the same ideas. This is the case even in text with very specific language such as cooking recipes. For example, for the ground truth *'Garnish with the remaining Wasabi and sliced green onions.'*, our method may predict *'Transfer to a serving bowl and garnish with reserved scallions.'*. For a human reader, this is half correct, especially since *'scallions'* and *'green onions'* are synonyms, yet this example would have only a BLEU1 score of 30.0, BLEU4 of 0.0 and METEOR of 11.00.

We report our results over the entire test set of the Recipe1M dataset [5] in Figures 6 (verbs and sentences), 7 (ingredients). We report scores of the predicted steps averaged over multiple recipes. Only those recipes which have at least j steps contribute to the average for step j. Compared to the recipes with exactly 9 steps, results over the entire test set are not significantly different in trends. Based on the ground truth, we observe that the majority of the ingredients occurs in the early and middle steps and decreases in the last steps, see Figure 7.

2.2. Human study

To assess the reliability of agreement between our human raters, we use Fleiss's kappa [2] measure. It is used to analyze how much the annotators agree in their decisions. High level of agreement (at most 1) indicates that the human rating study was reliable. Inter-rater agreement, measured via Fleiss's kappa [2] by aggregating across all rating tasks, is 0.43, which is statistically significant at $p < 0.05$.

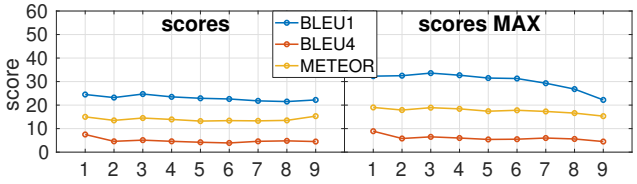

Figure 8: Comparison of the sentence scores versus the sentence scores computed based on the maximum match. Results reported over recipes with exactly 9 steps from Recipe1M dataset [5]. The x-axes indicate the step number being predicted in the recipe.

In our human study, we observe that even if the predicted step does not exactly match the ground truth, human raters still consider it possible for the future. Following this setting of the user study, we compute the sentence scores between the predicted sentence $\hat{s}_j$ and all future ground truth steps $\{s_j, s_{j+1}, s_{j+2}, s_{j+3}\}$ and select the step with the maximum score as our future match. We show the results for recipes with exactly 9 steps in Figure 8. The left

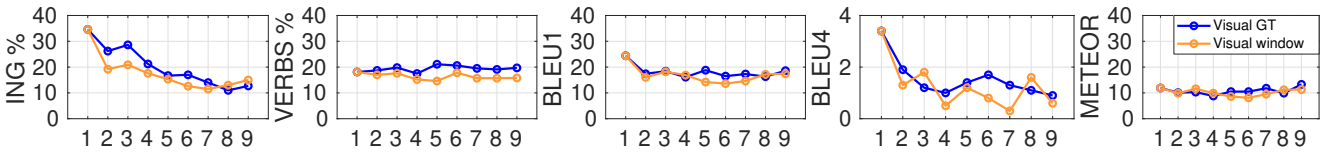

Figure 9: We compare the performance of our visual models for the recall of predicted ingredients and verbs, and sentence scores. Compared to using GT segments, the fixed windows show lower results but the results follow similar trends.

plot shows the standard scores between the predicted sentences and the ground truth. The right plot shows the scores computed based on the maximum future match.

We show some examples of predictions of our text-based method in Figures 13,15,18,19,20,16,14,17, along with the automated scores and human ratings.

2.3. Video Predictions

Window selection: We test two video segmentation settings for inference: one according to ground truth ("Ours Visual (GT)") and one based on fixed windows ("Ours Visual"). For "Ours Visual", we first partition the video until the last observation into fixed sized windows and sequentially feed these into our recipe RNN. Overall, our method is relatively robust to window size. We report results for different window sizes for YouCookII in Table 2 and for Tasty in Table 1.

Method	ING	verbs	BLEU1	BLEU4	METEOR
Ours Visual (window 70)	12.40	13.26	14.73	0.93	8.31
Ours Visual (window 90)	13.15	13.99	15.97	1.06	9.24
Ours Visual (window 110)	14.06	15.58	16.84	1.18	10.05
Ours Visual (window 130)	14.97	15.40	17.94	1.09	10.32
Ours Visual (window 150)	16.70	17.59	18.94	1.07	11.25
Ours Visual (window 170)	16.66	17.08	17.59	1.23	11.00
Ours Visual (window 190)	17.14	17.38	17.18	1.09	11.60
Ours Visual (window 210)	15.99	15.90	17.43	1.19	10.85
Ours Visual (window 230)	15.40	15.48	16.19	1.06	10.31

Table 1: Window size selection on the Tasty Videos dataset

Method	ING	verbs	BLEU1	BLEU4	METEOR
Ours Visual (window 30)	15.18	20.38	19.99	0.60	9.21
Ours Visual (window 50)	15.86	22.60	21.29	1.10	10.02
Ours Visual (window 70)	17.64	25.11	22.55	1.38	10.71
Ours Visual (window 90)	18.13	26.31	22.87	1.32	10.93
Ours Visual (window 110)	18.86	26.91	22.93	1.32	10.83
Ours Visual (window 130)	18.21	26.05	22.51	1.28	10.83
Ours Visual (window 210)	18.05	26.40	21.83	1.20	9.83

Table 2: Window size selection on the YouCookII dataset [6].

YouCookII dataset cross validation: To benchmark our model on the YouCookII dataset, we create a zero-shot setting using 4-fold cross validation. We create our splits based on distinct dishes. First set includes all videos from dish labels between 1 and 125, second set 126 and 222, third set 223 and 316 and fourth set 317 and 425.

YouCookII - more comparisons on supervised vs zero-shot performance: YouCook2 averages 22 videos per dish. We used 11 for testing; for supervised-training we use the other 11 for training, but exclude them for a comparable zero-shot scenario. Fig. 12 shows steady improvement when training with the dish-specific videos, indicating that the model is in fact learning and that more than 11 videos (current supervised setting) will further improve the supervised performance.

More results on the Tasty dataset: We compare the prediction performance of visual model with GT segments vs. window segment in Figure 9. Compared to using ground truth segments, the fixed window segments do not have a significant decrease in performance.

2.4. Ablation study

Since our method is modular, we conduct an ablation study to check the interchangeability of the sentence encoder on the Recipe1M dataset [5]. Instead of using our own sentence encoder, we represent the sentences using ST vectors trained on the Recipe1M dataset, as provided by the authors [5]. These vectors have been shown to perform well for their recipe retrieval. Our results, presented in Fig. 10 show that our sentence encoder performs on par with ST encodings. Moreover, our encoder, model and decoder can all be trained jointly and do not require a separate pre-training of a sentence autoencoder. In both cases the recipe RNN and sentence decoder have been tested with the same parameter settings. We also test how important the ingredients are as input for our method. We retrain our model without any ingredients using the same parameter settings. Results are shown in Figure 10; we see that ingredient information is very important for our method, especially in predicting the initial steps. However, in subsequent steps when 25%, 50% of the recipe steps are seen, the model's performance starts to improve as it receives more information.

2.5. Recipe Visualization

Our method can model recipes, as the output of the recipe RNN, especially after seeing all N steps, serves as a feature vector representing the entire recipe. For validating these features we conduct a recipe visualization experiment. We select recipes from the 9 most common recipe categories in the test set of the Recipe1M dataset [5] and encode them with our model by taking the final hidden output of the recipe RNN. As comparison, similar to [5], we take the mean of the ST vectors across the recipe steps. We visualize a two-dimensional representation computed using tSNE [3] of both recipe representations in Figure 11. We find that with our method, the recipes are better separated according to category.

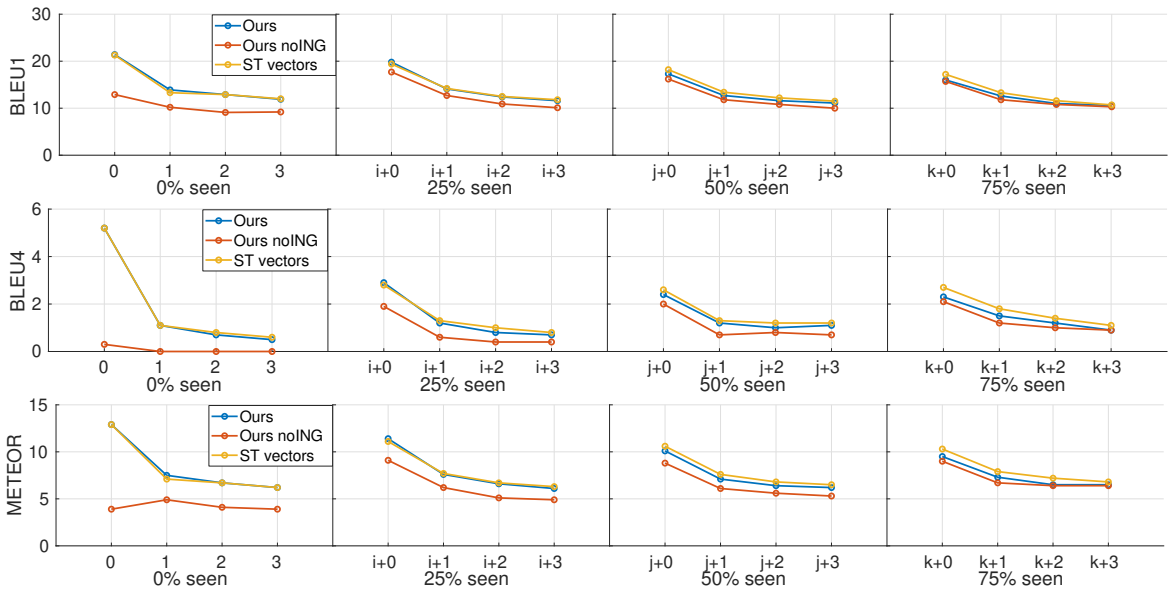

Figure 10: Ablation study to check the interchangeability of the sentence encoder and how important the ingredients are as input for our method, computed over the entire test set of the Recipe1M dataset. "$X\%$ seen" refers to the number of steps the model receives as input, while predicting the remaining $(100 - X)\%$.

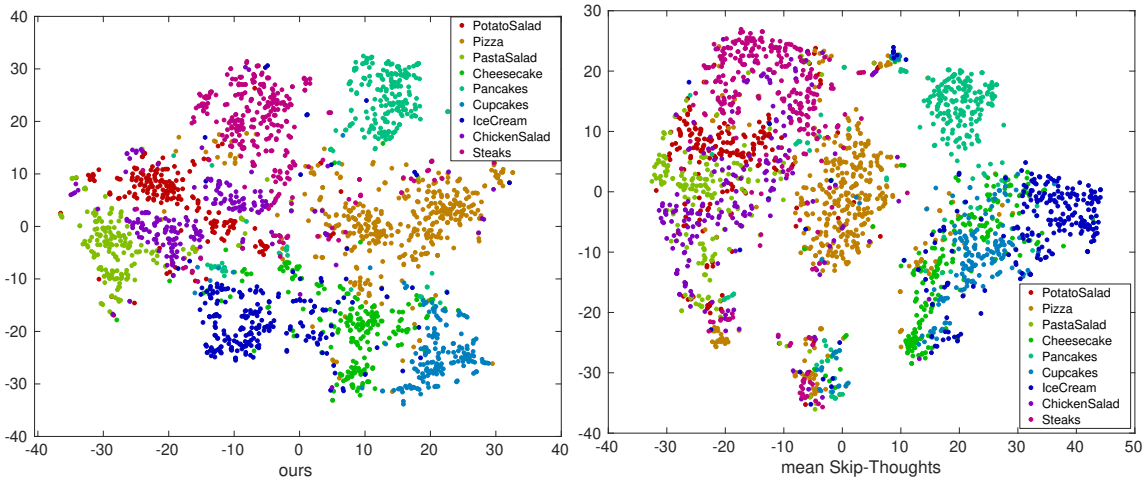

Figure 11: Recipe encoding visualization with tSNE [3] over a set of recipes from the 9 most common categories in the Recipe1M test set.

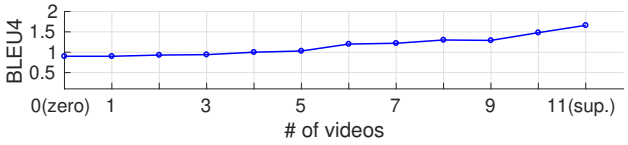

Figure 12: BLEU4 increases with # of training videos. 0/11 correspond to zero-shot/supervised settings; tested on YouCook2 set 1.

References

[1] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *ACL Workshop, 2005*.

[2] J.L. Fleiss et al. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382, 1971.

[3] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. volume 9, pages 2579–2605, 2008.

[4] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics, 2002*.

[5] Amaia Salvador, Nicholas Hynes, Yusuf Aytar, Javier Marin, Ferda Ofli, Ingmar Weber, and Antonio Torralba. Learning cross-modal embeddings for cooking recipes and food images. In *Conference on Computer Vision and Pattern Recognition (CVPR), 2017*.

[6] Luowei Zhou, Chenliang Xu, and Jason J Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI Conference on Artificial Intelligence, 2018*.

	ground truth	prediction	BLEU1	BLEU4	METEOR	HUMAN1	HUMAN2
ING	salt, pepper, onion, egg, bacon, ketchup, ground beef, Worcestershire sauce, parmesan cheese, hamburger buns, shredded cheddar cheese						
step1	Preheat a grill for high heat.	In a large bowl, combine the first six ingredients.	11.0	0.0	3.0	0.0	1.5
step2	In a large bowl, mix together the Cheddar cheese, Parmesan cheese, onion, egg, ketchup, Worcestershire sauce, salt and pepper.	In a large bowl, mix together the ground beef, onion, Worcestershire sauce, egg, salt	66.0	38.0	33.0	1.0	1.0
step3	Crumble in the ground beef, and mix together by hand.	Shape into 4 patties.	0.0	0.0	0.0	0.0	1.0
step4	Form into 6 patties, and wrap two slices of bacon around each one in both directions.	Form into 4 patties.	4.0	0.0	10.0	0.5	1.5
step5	Secure bacon with toothpicks.	Place on the grill, and cook for about 10 minutes, or until the bacon is cooked through	6.0	0.0	9.0	0.0	1.0
step6	Place patties on the grill, and cook for 5 minutes per side, or until well done.	Place on the grill, and cook for about 10 minutes, turning once.	48.0	32.0	22.0	1.5	1.5
step7	Remove toothpicks before serving on hamburger buns.	Serve on buns with lettuce, tomato, and ketchup.	25.0	0.0	14.0	0.5	1.5

Figure 13: Predictions of our text-based method for "Cheddar Bacon Wrapped Hamburgers" along with the automated scores and human ratings. step4 prediction is half correct. step7 is a plausible prediction.

	ground truth	prediction	BLEU1	BLEU4	METEOR	HUMAN1	HUMAN2
ING	milk, carrots, poultry seasoning, fresh ground black pepper, chicken bouillon cubes, celery ribs, boneless skinless chicken breasts						
step1	Place chicken in a slow cooker.	Place chicken in a large pot and cover with water.	40.0	26.0	29.0	0.5	1.0
step2	Heat broth in microwave ; dissolve bouillon in broth.	Add celery, carrots, and celery.	0.0	0.0	0.0	0.0	2.0
step3	Add next 4 ingredients to broth.	Pour over chicken.	0.0	0.0	0.0	0.0	1.5
step4	Pour over chicken.	Pour over chicken.	100	3.0	100	2.0	2.0
step5	Cover and cook on low for 6-8 hours, until chicken falls apart when poked with a fork.	Cover and cook on low for 8 hours.	28.0	23.0	26.0	1.5	1.5
step6	Combine buttermilk biscuit mix and milk, then drop spoonfuls over chicken to form dumplings.	Remove chicken from broth, cool and shred.	11.0	0.0	5.0	0.0	1.0
step7	Cover and cook on high for 35 minutes or until dumplings are done.	Cover and cook on low for 8 to 10 hours or until chicken is tender.	47.0	20.0	26.0	0.5	0.5

Figure 14: Predictions of our text-based method for "Slow Cooker Chicken and Dumplings" along with the automated scores and human ratings. step2 prediction is a plausible future step. step5 is correct.

	ground truth	prediction	BLEU1	BLEU4	METEOR	HUMAN1	HUMAN2
ING	pineapple, strawberries, oranges, flaked coconut, fresh mint leaves, vanilla yogurt, kiwi fruits						
step1	Quarter pineapple lengthwise ; remove core.	In a large bowl, combine the strawberries, kiwi, oranges, and pineapple.	9.0	0.0	7.0	0.0	1.0
step2	Cut crosswise into small chunks.	Cut into 1/2-inch cubes.	39.0	0.0	12.0	1.0	1.0
step3	Place in large serving bowl.	In a large bowl, combine fruit, kiwi, and pineapple.	33.0	0.0	20.0	0.0	1.0
step4	Add clementine segments, strawberries, kiwifruit, and coconut ; gently toss.	Add fruit and nuts.	14.0	0.0	7.0	0.5	0.5
step5	Spoon into dessert glasses.	Serve immediately or store in refrigerator up to 3 days.	0.0	0.0	0.0	0.0	0.5
step6	Top with a dollop of vanilla yogurt or sweetened sour cream.	Garnish with orange slices and mint.	7.0	0.0	2.0	0.0	1.0
step7	Garnish with mint sprigs if desired.	Garnish with orange slices and mint.	50.0	0.0	22.0	1.0	1.0

Figure 15: Predictions of our text-based method for "Ambrosia Fruit Salad" along with the automated scores and human ratings. step1, step3 and step6 are plausible future step predictions.

	ground truth	prediction	BLEU1	BLEU4	METEOR	HUMAN1	HUMAN2
ING	butter, cheese, eggs, salt and pepper						
step1	Whisk the eggs till light and fluffy.	Preheat oven to 350 degrees.	0.0	0.0	0.0	0.0	0.0
step2	Add cheese, salt and pepper.	Add the cheese and season with salt and pepper.	56.0	0.0	43.0	2.0	2.0
step3	Mix well.	Melt butter in a pan.	0.0	0.0	0.0	0.0	1.5
step4	Heat a 7 inch non-stick skillet and add butter.	Melt butter in a pan.	18.0	0.0	11.0	1.0	1.0
step5	As the butter begins to smoke, lower heat and pour the beaten egg mixture.	Pour in egg mixture.	6.0	0.0	14.0	1.0	1.0
step6	Cook covered for a minute.	Let cook until eggs are set.	17.0	0.0	10.0	1.5	1.5
step7	Fold over and serve immediately.	Turn the heat down to low and cook for about 5 minutes.	8.0	0.0	3.0	0.0	1.0

Figure 16: Predictions of our text-based method for "Cheese Omelette" along with the automated scores and human ratings. step5 is a correct prediction. step7 is a plausible prediction.

	ground truth	prediction	BLEU1	BLEU4	METEOR	HUMAN1	HUMAN2
ING	sugar, celery, black olives, chopped tomatoes, green olives						
step1	Boil your noodles as directed on box.	Mix all ingredients together.	0.0	0.0	0.0	0.0	0.5
step2	Do not overcook!	Drain and rinse with cold water.	0.0	0.0	0.0	0.0	2.0
step3	Drain.	Drain and rinse with cold water.	17.0	0.0	26.0	1.5	1.5
step4	Add veggies to pasta in large bowl.	In a large bowl, mix together the noodles, celery, tomatoes, olives, and sugar.	23.0	0.0	16.0	1.5	1.5
step5	Add Italian dressing and Splenda or sugar and ground pepper.	Add all other ingredients.	6.0	0.0	6.0	1.5	1.5
step6	Mix well, chill and enjoy!	Mix well.	22.0	0.0	24.0	1.5	1.5
step7	Add more dressing the next day as needed, if put in the fridge overnight!	I like to add a little bit of olive oil to the salad and I add a little more	16.0	0.0	6.0	0.5	1.0

Figure 17: Predictions of our text-based method for "My Mom's Summer Pasta Salad" along with the automated scores and human ratings. step2 is a plausible future step. step4 is correct as veggies refer to celery tomatoes and olives. step7 is a plausible suggestion even though it does not match the GT step.

	ground truth	prediction	BLEU1	BLEU4	METEOR	HUMAN1	HUMAN2
ING	salt and pepper, parmesan cheese, garlic cloves, extra virgin olive oil, diced tomatoes, fusilli, pesto sauce, fresh basil leaves						
step1	Heat oil in a medium skillet over medium heat.	Heat the oil in a large skillet over medium heat.	80.0	39.0	43.0	2.0	2.0
step2	Saute garlic for 2 minutes, then add tomatoes and simmer for 10 minutes.	Add garlic and cook until fragrant, about 30 seconds.	21.0	0.0	9.0	0.5	1.5
step3	Meanwhile, bring a large pot of lightly salted water to a boil.	Add pesto and cook for 1 minute.	0.0	0.0	4.0	0.0	1.5
step4	Add pasta and cook for 8 to 10 minutes or until al dente ; drain and return to pot.	Add pasta and cook for 8 to 10 minutes or until al dente ; drain.	75.0	75.0	54.0	2.0	2.0
step5	Season tomato sauce with salt and pepper ; stir in pesto.	Add pesto and basil to pasta and toss to combine.	20.0	0.0	7.0	0.5	1.5
step6	Pour sauce over hot cooked pasta and mix well.	Pour sauce over pasta and toss to coat.	55.0	0.0	25.0	1.0	1.0
step7	Sprinkle with cheese, garnish with basil leaves and serve immediately.	Serve immediately.	2.0	0.0	14.0	1.0	1.0

Figure 18: Predictions of our text-based method for "Easy Fusilli with Tomato Pesto Sauce" along with the automated scores and human ratings. step3 is a plausible future prediction. step5 is half correct.

	ground truth	prediction	BLEU1	BLEU4	METEOR	HUMAN1	HUMAN2
ING	salt, butter, cheese, cream cheese, scallions, cauliflower, black pepper, bacon bits, whole milk						
step1	Preheat oven to 350	Preheat oven to 350 degrees F.	67.0	51.0	50.0	2.0	2.0
step2	Chop cauliflower and boil in salted water for about 3-4 mins until a little tender	Place cauliflower in a large pot and cover with water.	30.0	0.0	10.0	1.0	1.0
step3	Strain cauliflower and mash with a potato masher.	Add the butter, salt, and pepper, and stir to combine.	10.0	0.0	2.0	0.0	1.5
step4	Place in bowl and mix all ingredients	Melt butter in a large skillet over medium heat.	11.0	0.0	3.0	0.0	1.0
step5	Add ingredients to a buttered 8x8 dish and top with extra cheese and bacon.	Pour mixture into a baking dish and bake for 20 minutes	21.0	0.0	7.0	0.5	1.5
step6	Bake for 30 mins until nice and bubbly.	Bake for 30 minutes.	27.0	0.0	28.0	2.0	2.0
step7	Enjoy!	Serve with a side of buttered toast.	0.0	0.0	0.0	0.5	1.5

Figure 19: Predictions of our text-based method for "Baked Loaded Cauliflower" along with the automated scores and human ratings. step5 is half correct.

	ground truth	prediction	BLEU1	BLEU4	METEOR	HUMAN1	HUMAN2
ING	salt, butter, potatoes, sour cream, green onions, sharp cheddar cheese						
step1	Peel, quarter, and boil potatoes until barely tender.	Preheat oven to 350 degrees.	0.0	0.0	0.0	0.0	2.0
step2	Drain, cool and grate on large side of grater.	Melt butter in a large skillet over medium heat.	11.0	0.0	6.0	0.0	2.0
step3	Place in a large bowl.	In a large skillet, melt butter over medium heat.	33.0	0.0	21.0	0.0	2.0
step4	Gently mix in all other ingredients, except butter.	Add remaining ingredients and mix well.	24.0	0.0	17.0	1.0	1.0
step5	Place mixture into a greased shallow 9x9 pan.	Spoon mixture into a lightly greased 9 x 13 inch baking dish.	33.0	0.0	18.0	1.5	1.5
step6	Top with melted butter.	Bake at 350 degrees for 30 minutes.	0.0	0.0	0.0	0.0	1.5
step7	Bake at 400 degrees for 30 minutes or until golden.	Bake at 350 degrees for 1 hour.	37.0	0.0	21.0	1.5	1.5

Figure 20: Predictions of our text-based method for "Green Onion Potato Casserole" along with the automated scores and human ratings. step1 is a plausible prediction as the oven will be used for baking. step6 is a plausible future step.

ING: salt, butter, flour, garlic, olive oil, cloves, warm water, bread flour, fresh parsley, dry yeast

1 **GT:** For the dough, mix together the **flour**, **yeast**,and **salt**, and make a well in the middle.
PRED: In a small bowl, combine the **yeast**, water and **salt**.

2 **GT:** Pour the water and 1 teaspoon of olive oil into the well and **mix together**.
PRED: Add the yeast mixture and **mix** until the dough comes **together**.

3 **GT: Knead** for about 7-10 minutes until the dough is springy.
PRED: Knead the dough for a few minutes, then cover and leave to rise for 1 hour.

4 **GT:** Place the dough in the bowl, cover, leave to rise in a warm place for 1 hour.
PRED: Knead the dough for 5 minutes, then knead for a few minutes until smooth and elastic.

5 **GT:** Tear the dough into 9 separate balls and then **roll** into long tubes, folding over to make a knot.
PRED: Once the dough has risen, **roll** it out on a floured surface to a thickness of about 1/4.

6 **GT: Place** the knots on a **baking tray**.
PRED: Place on a **baking tray** lined with parchment paper.

7 **GT: Bake** in a 350°F (180°C) oven for 10 minutes.
PRED: Bake for 20 minutes, or until the bread is golden brown.

8 **GT:** Prepare the garlic butter by mixing the butter, garlic, and parsley.
PRED: Serve hot.

9 **GT:** Coat the knots with the **butter mixture** and bake for another 5 minutes
PRED: Brush the garlic **butter mixture** over the bread.

10 Enjoy! Enjoy!

Figure 21: Next step prediction of our visual model for "Garlic Knots". The blue sentences are our model's predictions. After baking, in step7, our model predicts that the dish should be served, but after visually seeing the butter parsley mixture in step8, it correctly predicts that the knots should be brushed in step9. Note that our model predicts the next steps before seeing these segments!

ING: salt, pepper, onion, milk, egg, meatballs, ground beef, Worcestershire sauce, parmesan cheese, canola oil, beef broth, fresh parsley, seasoned bread crumbs

1 **GT:** In a large bowl, **combine** the **ground beef**, **breadcrumbs**, **onion**, **egg**, salt, and pepper, mixing until evenly combined.
PRED: In a large bowl, **combine** the ground **beef**, **bread crumbs**, milk, **egg**, parsley, **onion**

2 **GT:** Heat the canola oil in a pot over medium-high heat.
PRED: Pour mixture into a loaf pan and bake for about 1 hour.

3 **GT:** Place the **meatballs** into the pot, **cooking** for one minute.
PRED: Add the **meatballs** to the pan and **cook** for about 5 minutes on each side.

4 **GT:** Flip the meatballs.
PRED: In a medium bowl, combine the meatballs, sauce, and parsley.

5 **GT:** Add the beef broth, milk, salt, pepper, and Worcestershire sauce and give it a stir.
PRED: Transfer the meatballs to a baking sheet lined with parchment paper.

6 **GT:** Bring the liquid to a boil, then add the egg noodles..
PRED: Pour the meatballs into the baking dish.

7 **GT: Stir** constantly until the pasta is cooked and the liquid has reduced to a sauce that coats the noodles, about seven to eight minutes..
PRED: Add the meatballs to the sauce and **stir** to combine.

8 **GT:** Add the parmesan and the parsley, stirring until the cheese is melted.
PRED: Enjoy!

9 Enjoy! Enjoy!

Figure 22: Next step prediction of our visual model for "One-pot Swedish Meatball Pasta". The blue sentences are our model's predictions. Our prediction for step3 matches the GT step4. Our model's prediction for step4 is somehow plausible as a future step as the GT in step5 suggest mixing the sauce and meatballs. Note that our model predicts the next steps before seeing these segments!

ING: chocolate, honey, almonds, raisins, candy, peanuts, rolled oats, nut butter

1 **GT:** Melt peanut butter and honey together in the microwave, stirring every 15 seconds.
PRED: In a large bowl, combine the oats chocolate chips, **honey**, and **peanut butter**

2 **GT:** Add **oats**, almonds, **peanuts**, and raisins into the bowl and mix until well coated.
PRED: Stir in **oats** and **peanuts**.

3 **GT:** Add chocolate candies and mix just until combined.
PRED: Press into a greased 9x9-inch pan.

4 **GT:** Line a baking dish with parchment paper and pour the cooled mixture into the pan.
PRED: Using a small, sharp knife, cut into squares.

5 **GT:** Cool in the refrigerator for at least 2 hours.
PRED: Enjoy !

6 Enjoy! Enjoy!

Figure 23: Next step prediction of our visual model for "Trail Mix Bars". The blue sentences are our model's predictions. Our model's prediction for step3 is a plausible future step prediction as the mixture is placed in a pan in step4. Note that our model predicts the next steps before seeing these segments!

ING: salt, pepper, olive oil, tomatoes, green beans, chicken thighs, basil pesto

1 **GT:** In a large pan, heat olive oil and add chicken thighs.
PRED: Preheat oven.

2 **GT: Season with salt and pepper**.
PRED: Season with salt and pepper.

3 **GT:** When the chicken is completely **cooked through**, remove from pan**.**
PRED: Cook for about 5 minutes on each side, or until **cooked through.**

4 **GT:** Slice into strips, and set aside.
PRED: Add the green beans to the pan and cook for 2 minutes.

5 **GT:** Add green beans and cook until crisp tender.
PRED: Place the chicken thighs on top of the pesto.

6 **GT:** Return the chicken strips to the pan, then **add tomatoes** and **pesto**.
PRED: Add the **pesto** and **tomatoes**.

7 **GT:** Stir until fully incorporated.
PRED: Pour the remaining sauce over the chicken and vegetables**.**

8 **GT:** Divide into 4 food storage containers and store in the refrigerator.
PRED: Enjoy!

9 Or serve immediately! Enjoy!

Figure 24: Next step prediction of our visual model for "Weekday Meal-prep Pesto Chicken and Veggies". The blue sentences are our model's predictions. Our model's prediction for step4 is a plausible future step prediction as it happens in step5. Our model's predictions for step8 and step9 are plausible recommendations. Note that our model predicts the next steps before seeing these segments!