Recipe2Video: Synthesizing Personalized Videos from Recipe Texts
Supplementary Material

Prateksha Udhayanan[1], Suryateja BV[2]*, Parth Laturia[3]*,
Dev Chauhan[4]*, Darshan Khandelwal[5]*, Stefano Petrangeli[1], and Balaji Vasan Srinivasan[1]
[1]Adobe Research; [2]Avanti Fellows; [3]Morgan Stanley; [4]Graviton Research Capital LLP; [5]Goldman Sachs
udhayana@adobe.com, suryateja@avantifellows.org, dev.chauhan@gravitontrading.com
parthlaturia@gmail.com, darshankhandelwal1218@gmail.com, {petrange,balsrini}@adobe.com

1. Introduction

The supplementary material is organized as follows:

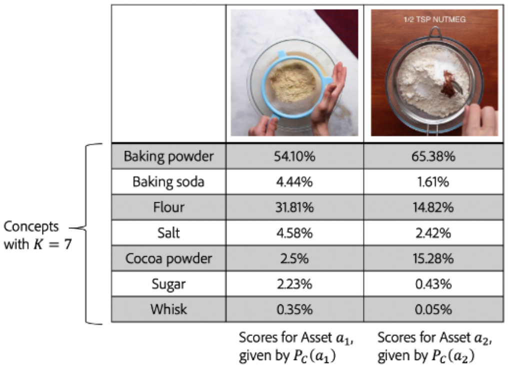
1. Example-based explanation of different components of the Recipe2Video architecture
2. Qualitative examples
3. Details on the human evaluation process
4. Demo video

2. Recipe2Video: System Architecture

To build our Recipe2Video system, we adopt the framework of diagrammatic mode from modern multimodal theory [2], which offers a top-down approach for designing effective multimedia content: (i) decide a communicative goal of the content, also called discourse semantics; (ii) find expressive resources to meet the chosen communicative goals. Food recipe texts are one of the few domains that offer data [3, 1] connecting multimedia and communicative goals. We therefore convert recipe texts to videos with the goal of enhancing consumption experiences leveraging these data [3, 1].

2.1. Ranking assets and their combinations

Figure 1 shows an example of computing **information coverage** scores to rank two retrieved assets. We get $K = 7$ key phrases and compute KL divergence scores of assets a_1 and a_2 to obtain 0.0028 and 0.0038 respectively. Thus, we choose asset a_1 over asset a_2. Visually, we can see that asset a_2 contains cocoa powder in the bowl and hence gets a higher score for the "Cocoa Powder" concept (15.28%) whereas asset a_1 gets a low score (2.5%) for the same key phrase. However, asset a_1 has a greater score in most other key phrases such as "Baking Soda", "Flour", "Sugar", thus leading to a lower KL divergence value.


1/2 TSP NUTMEG
Concepts with $K = 7$

	Scores for Asset a_1, given by $P_C(a_1)$	Scores for Asset a_2, given by $P_C(a_2)$
Baking powder	54.10%	65.38%
Baking soda	4.44%	1.61%
Flour	31.81%	14.82%
Salt	4.58%	2.42%
Cocoa powder	2.5%	15.28%
Sugar	2.23%	0.43%
Whisk	0.35%	0.05%

Figure 1. An example of using Information Coverage for ranking two assets. These assets are retrieved for the instruction: *In a large bowl, whisk together the sugar, flour, cocoa powder, baking powder, baking soda, and salt*

.

As mentioned before, we score the **temporal assets** based on 3 questions: (1) Does the image show how to prepare before carrying out the instruction? (2) Does the image show results of the action described in the instruction? (3) Does the image depict an action in progress described in the instruction? The characterization of the temporal aspects into three categories allows us to synthesize video according to specific user preferences. For example, a consumer looking for a succinct summary of the actions might be better served by optimizing the assets for the third question above, while someone who is preparing for a procedure can be better served by optimizing for the first question above. Figure 2 shows a single asset (a) and a 3-asset (b) example that are likely to get chosen for different video variants.

3. Qualitative Examples

Figure 3 shows a set of frames extracted from the elaborate video variant synthesized by Recipe2Video system for a *chocolate cake* recipe. Precise illustrations of canola

*Work done while at Adobe Research

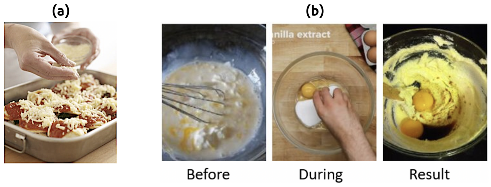
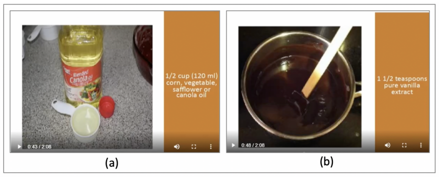
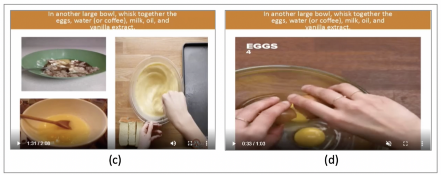


Figure 2. Part (a) demonstrates a single asset that illustrates the instruction: *Top with another layer of ravioli and the remaining sauce not all the ravioli may be needed. Sprinkle with the Parmesan*. The image depicts both the action and the result of the action. An asset like this is a potential candidate of being chosen for the succinct variant. Part (b) demonstrates a 3-asset combination chosen for instruction: *In a large bowl, whisk together the eggs, water, milk, oil, and vanilla extract*. This combination is very likely to be chosen for the elaborate variant.


1/2 cup (120 ml) corn, vegetable, safflower or canola oil

1 1/2 teaspoons pure vanilla extract

Figure 3. Frames extracted from the elaborate video variant synthesized by our system. Both frames (a, b) illustrate instructions which could potentially help a novice user understand the prerequisites of making Chocolate Cake recipe.


In another large bowl, whisk together the eggs, water (or coffee), milk, oil, and vanilla extract.

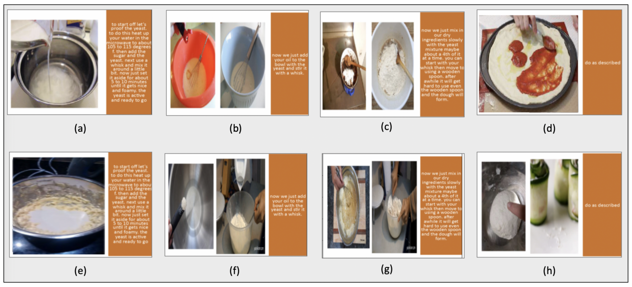
In another large bowl, whisk together the eggs, water (or coffee), milk, oil, and vanilla extract.

Figure 4. Two frames corresponding to the same instruction but from different variants synthesized by our system. Frame (c) is extracted from the elaborate variant; Frame (d) is extracted from the succinct variant.

oil and vanilla extract are displayed in (a) and (b) respectively, which could potentially help a novice user. Figure 4 compares the chosen final assets for the same instruction text in the elaborate (c) and succinct (d) variants. As evident from frame (c), the assets contain the state of the ingredients (eggs, water, oil) for all three temporal aspects (before, during, after) which aids in providing a holistic understanding of procedure to the user. This is possible due to the presence of our ranking module that elevates the combination of assets with strong temporal aspects. In frame (d) corresponding to the succinct variant, only one asset (video clip) is chosen that provides a quicker and concise representation of the given instruction. Frames (a, b, c) can be found at timestamps (0:43, 0.48, 1:31) respectively in the video titled *Recipe2Video_Elaborate_QE1.mp4*, Frame (d) can be found at timestamp (0:33) in the video *Recipe2Video_Succinct_QE1.mp4* in the following link - `https://bit.ly/3s11wp3`.

Figure 5 contains frames (a-d) extracted from a succinct video synthesized by Recipe2Video, while the frames (e-h) are extracted from a video synthesized by the *Audiovisual Slideshows* baseline for a *Homemade Pizza Dough* recipe. The videoclip asset in frame (a) shows "hot water" being added that results in "nice and foamy" yeast, as indicated by the instruction text. However, the equivalent frame (e) from the baseline does not convey the entire information, pointing to the strength of our method in selecting appropriate multimodal assets.

Comparing frame (b) with frame (f), we see that our method retrieves much better assets corresponding to the "stir with whisk" instruction. On comparing frame (c) and (g), we note that Recipe2Video is able to retrieve appropriate assets containing both "yeast" and "wooden spoon" for whisking, while the baseline incorrectly contains only "whisk". This shows the strong visual and textual relevance of our assets owing to the proposed retrieval and ranking module. Certain texts contain no semantic information and act as connectors between two steps. 'Do as described' (d & h) is one such example in Figure 5. While (d) contains a meaningful image corresponding to a topping being added to the pizza dough, frame (h) contains two completely unrelated assets to the context of the instruction. This shows the strength of our Viterbi decoding step that leverages inter-frame semantics to achieve overall coherence. Frames (a, b, c, d) can be found at timestamps (0:10, 0:15, 0:25, 1:05) respectively in the video *Recipe2Video_Succinct_QE2.mp4*, Frames (e, f, g, h) can be found at timestamps (0:18, 0:32, 0:49, 1:29) in the video *Baseline_Succinct_QE2.mp4* in the following link - `https://bit.ly/3s11wp3`.


to start off let's proof the yeast. to do this heat up your water in the microwave to about 105 to 115 degrees f. then add the sugar and the yeast. next use a whisk and mix it around a little bit. now just set it aside for about 5 to 10 minutes until it gets nice and foamy. the yeast is active and ready to go

now we just add your oil to the bowl with the yeast and stir it with a whisk.

now we just mix in our dry ingredients slowly with the yeast mixture maybe about a 4th of it at a time. you can start with your whisk then move to using a wooden spoon. after awhile it will get hard to use even the wooden spoon and the dough will form.

do as described

to start off let's proof the yeast. to do this heat up your water in the microwave to abou 105 to 115 degrees f. then add the sugar and the yeast. next use a whisk and mix it around a little bit. now just set it aside for about 5 to 10 minutes until it gets nice and foamy. the yeast is active and ready to go

now we just add your oil to the bowl with the yeast and stir it with a whisk.

now we just mix in our dry ingredients slowly with the yeast mixture maybe about a 4th of it at a time. you can start with your whisk then move to using a wooden spoon. after awhile it will get hard to use even the wooden spoon and the dough will form.

do as described

Figure 5. Comparison of frames extracted from videos synthesized by our Recipe2video method and Audiovisual Slideshows method. Frames (a-d) correspond to our system; Frames (e-h) correspond to Audiovisual Slideshows baseline. Each column represents the same instruction but the assets in our video are more informative, coherent, and thereby enhancing consumer experience as opposed to the baseline video.

4. Human Evaluation

For all MTurk surveys, we set the annotator prerequisites as "MTurk Masters" located in the United States having an

[Sanity Check, Task; 1, 2, 3] How many instructions/steps are present in this recipe?
● 4 or 5 ● 6 or 7 ● 8 or 9 ● 10 or 11

[Sanity Check, Task; 1, 2, 3] What could be an appropriate name/title for the recipe shown above?
Respondent chooses one of the four options which contain one correct answer along with three randomly sampled names

[Sanity Check, Task; 1, 2, 3] Which of the following ingredients might have been used in the recipe?
Respondent chooses one of the four options which contain one correct answer along with three randomly sampled names

[Enjoyable; 1, 2] How enjoyable or boring did you find going through the recipe?
● Very boring ● A little boring ● A little enjoyable ● Very enjoyable

[Retainable; 1, 2] How much of the recipe can you remember now without looking at it again?
● Cannot remember anything ● Can remember very little ● Can remember some of it ● Can remember most

[Jarring; 1, 2] How often did you find yourself going back to previous steps in the text instructions list to understand the recipe?
● Never ● Once ● Sometimes ● Almost after every step

[Task; 1, 2] Select the incoherent image in the following set of images. An image is incoherent if it is not related to the recipe.
● Image-1 ● Image-2 ● Image-3 ● Image-4

[Intra-Coherence; 2] Were the images/clips within the frames related to the text shown in the frame?
● Not at all ● A little ● Somewhat ● A lot

[Inter-Coherence; 2] Do you think the steps displayed in the video along with images and clips were sensible and followed the right order?
● Made no sense at all ● Mostly didn't make any sense ● Somewhat made sense ● Made perfect sense

[Relevance; 3] Were the images/clips in the video related to the recipe text?
● Not at all ● A little ● Somewhat ● A lot

[Self-correction; 3] Suppose you made a mistake while preparing this recipe in your kitchen. What would you opt to view again to correct yourself? ● Video ● Text ● None

[Feedback; 1, 2, 3] Please provide a feedback on this survey

Table 1. Here are a few sample questions that were presented to the annotators for human evaluations.

approval rate $\geq 95\%$ and at least 100 annotations approved in the past. Using dataset statistics (average word count, average step count and average video duration) and several pilot runs on MTurk, we estimate the mean time taken to complete our surveys. We use this information to decide a budget for our experiments and the number of videos to be annotated, and pay our annotators at 12/hour.

Table 1 lists a set of example questions that were presented to the annotators for human evaluation. We also indicate the metrics and experiments each question correspond to. For instance, **[Enjoyable; 1, 2]** means that this question is used in experiments (1) and (2) to gauge the enjoyability of the displayed modality.

Figure 6 shows the average number of respondents who prefer to consume actions via visuals. This further verifies our hypothesis that the consumption of a procedural text is better carried out via visuals than text.

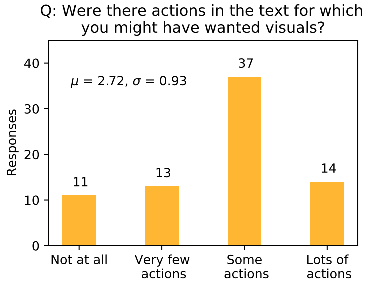


Figure 6. Average preference of respondents on a Likert Scale, when they were shown recipe texts in Expt (1) and asked if there were actions where they might have preferred visuals. Note that the total number of responses is $N \times 5 = 75$, since we evaluated on $N = 15$ RecipeQA texts.

As described before, in Expt (3), we ask the respondents to opt for their preferred modality (Video/Text/None) of consumption under different scenarios. Figure 7 shows the aggregated responses across both RecipeQA (23 recipes) and Tasty Videos (25 recipes) datasets for different scenarios. To capture the respondent bias of opting for "Video" regardless of the situation, we calculate the average across all scenarios and variants for a particular system and display it as a horizontal line (125.2 for Recipe2Video, 55.7 for Baseline). Respondents generally opt for our elaborate

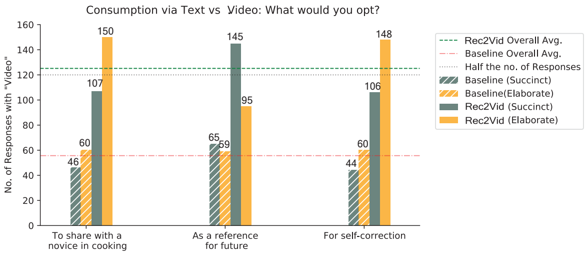


Figure 7. The bar graphs show the number of respondents who chose ”Video” for each of the questions listed in the x-axis, aggregated over both datasets. Note that each bar here corresponds to $(23+25) \times 5 = 240$ responses, since we solicit five responses for each (text, video) task pair in Expt (3).

video variant over the procedural text when they would need to share it with a novice who is new to cooking. Similarly, respondents opt for our succinct variant to use as a reference for the future. Also, there is a significant gain $(+24.8)$ in opting for the elaborate variant (150) over the average (125.2) of opting for videos of Recipe2Video system. Similarly, succinct variants offer a gain of 19.8 over the average. These results show that the synthesized variants meet the intended communicative goals.

5. Demo Video

A working demo video of our end-to-end system Recipe2Video can be found in the following link - `https://bit.ly/3s11wp3`. The video file is titled *Recipe2Video_Demo.mp4*. The video contains a demonstration of the working of the system - it takes an input document and generates 2 variants of videos catering to different user preferences. The demo also explains the working of the ranking modules with several examples.

References

[1] Malihe Alikhani, Sreyasi Nag Chowdhury, Gerard de Melo, and Matthew Stone. CITE: A corpus of image-text discourse relations. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 570–575. Association for Computational Linguistics, 2019.

[2] Tuomo Hiippala and John A Bateman. Semiotically-grounded distant viewing of diagrams: insights from two multimodal corpora. *arXiv preprint arXiv:2103.04692*, 2021.

[3] Semih Yagcioglu, Aykut Erdem, Erkut Erdem, and Nazli Ikizler-Cinbis. Recipeqa: A challenge dataset for multimodal comprehension of cooking recipes. *CoRR*, abs/1809.00812, 2018.