Seq2Time: Sequential Knowledge Transfer for Video LLM Temporal Grounding (Supplementary Materials)

Andong Deng[1,2,*], Zhongpai Gao[2,†], Anwesa Choudhuri[2], Benjamin Planche[2], Meng Zheng[2], Bin Wang[2,3], Terrence Chen[2], Chen Chen[1], Ziyan Wu[2]

[1]University of Central Florida, [2]United Imaging Intelligence, [3]Northwestern University

1. Data Preparation

1.1. Image Sequence

Data Source. To construct an image sequence dataset that provides high-quality visual-text alignment, we select LLaVA-ReCap [4] as the primary source. Built on the robust vision-language model LLaVA-NeXT-34B [5], LLaVA-ReCap contains over 3.7 million image-text pairs. Its captions are detailed, providing a rich description of the visual content, as exemplified in Figure 1. These captions enable strong correspondence between visual elements and language, serving as a reliable basis for generating high-quality sequential supervision in our training. This alignment is critical for training models to understand and reason over temporally structured visual data.

Instruction Template Design. To maximize the diversity and effectiveness of the training data, we design multiple instruction templates for each pretext task using GPT-4o [6]. These templates include variations in both questions and answers to ensure robustness in model learning. Figures 2, 3, 4, 5, and 6 illustrate examples of these templates for different tasks, such as Image Index Grounding (IIG), Indexed Image Captioning (IIC), and Adjacent Location Reasoning (ALR). This variety enhances the model's ability to generalize across different types of instructions and strengthens its visual-language understanding capabilities.

1.2. Clip Sequence

Data Source. We utilize the Kinetics-700 dataset [1], a large-scale video understanding dataset comprising approximately 650,000 ten-second video clips annotated with 700 distinct human action classes. The dataset offers diverse and naturalistic videos capturing a wide range of everyday activities. For our Seq2Time, we focus on the training split, which contains about 550,000 clips, ensuring a broad spectrum of visual and temporal information for model training.

Clip Captioning with LongVA. To generate textual descriptions for the video clips, we employ LongVA [7], a state-of-the-art video captioning model. LongVA generates captions based on the visual content and the associated action label, as demonstrated in Figure 7. For instance, it accurately captures detailed descriptions like "`A player in an orange jersey is leaping towards the hoop...`" for basketball action clips. These captions provide contextual insights into the actions and environments depicted in the videos, enriching the clip sequence data with semantic annotations.

Despite its strengths, LongVA occasionally produces hallucinated descriptions, as illustrated in Figure 8. For example, a video showing an adult and a baby reading by a table might be misinterpreted as "`a group of adults and babies...`" Such hallucinations can potentially introduce noise into the dataset, affecting the reliability of the generated sequences. This highlights a trade-off between the scalability of automatic captioning and the quality of annotations.

2. Visualization

In this section, we present qualitative results to demonstrate the effectiveness of our Seq2Time. Compared with the baseline model TimeChat, Seq2Time significantly improves upon repetitive text patterns, temporal localization accuracy, and the precision of event descriptions.

In Figure 9, we analyze a video illustrating the instructions for making pancakes. TimeChat struggles to recognize ingredients in certain steps, providing incomplete descriptions such as "`a bowl of eggs`" and "`a bowl of flour,`" and its temporal predictions are imprecise, resulting in a fragmented understanding of the cooking process. In contrast, Seq2Time generates accurate and comprehensive event descriptions with precise temporal annotations, such as predicting "`The next step is to mix the ingredients together in a bowl`" as occurring between 25.0 to 61.4 seconds, closely aligning with the ground truth annotation: "`mix flour,`

sugar, baking powder, and salt together in a bowl" from 25 to 61 seconds. This comparison highlights Seq2Time's ability to capture finer-grained details, maintain alignment with the video's temporal structure, and robustly capture the sequential flow of actions, effectively reducing redundancies and hallucinations often present in TimeChat's outputs. These improvements underscore Seq2Time's effectiveness in enhancing temporal grounding and understanding for long videos.

In Figure 10, a video illustrating the steps for cooking eggs is analyzed. TimeChat exhibits several issues, including generating repetitive text such as "She pours some into the bowl and stirs it" across multiple timestamps and misrecognizing objects, as seen in the caption "She sets the bowl on the stove," where the "bowl" should be the "pan." In contrast, Seq2Time minimizes these repetitive patterns and provides a more coherent description of the cooking process. Although Seq2Time incorrectly predicts the timestamp for the event "She stirs the mixture again," it successfully captures the overall sequence of actions with detailed and contextually appropriate descriptions. This example highlights Seq2Time's enhanced capability to handle complex sequences with improved temporal and semantic accuracy compared to TimeChat.

In Figure 11, the video illustrates the steps for making tacos. TimeChat fails to recognize any of the correct cooking steps, producing entirely incorrect predictions and hallucinating that the video involves baking in an oven. In contrast, our Seq2Time provides general but accurate descriptions of the steps, despite omitting some specific ingredient details. For example, Seq2Time predicts "The man adds more ingredients to the dish and continues to stir it together" between 148.1 and 259.2 seconds. While this aligns with the overall sequence, the ground truth includes finer-grained annotations, such as adding specific ingredients like cumin powder and beef.

Figure 12 presents a video of two girls preparing tofu soup. Our Seq2Time successfully captures key cooking steps, such as describing "They mix the ingredients in a pot and stir it" from 136.1 to 211.4 seconds. This aligns closely with the ground truth annotation, "add the tofu chunks and dissolve miso paste in the soup," which spans 142 to 184 seconds. In contrast, TimeChat only captures the initial scene of the video, which is less relevant to the cooking process, failing to identify any meaningful steps.

Figure 13 shows a video depicting a woman making cakes. TimeChat fails to generate relevant captions, producing incorrect descriptions, such as "a young man baking something." While Seq2Time accurately identifies basic events within the video, providing correct and relevant captions that align with the overall context.

Finally, Figure 14 shows a video of a woman preparing a salad in the kitchen. Both TimeChat and Seq2Time manage to produce correct captions for this video. However, neither model provides detailed step-by-step instructions. Instead, they capture only general actions, such as "woman talks" or "show the different ingredients," overlooking specific details of the cooking steps. This indicates an area where further refinement of both models could enhance their performance in capturing detailed procedural actions.

These examples collectively demonstrate the superiority of our Seq2Time in reducing hallucinations, improving temporal alignment, and providing more accurate descriptions compared to TimeChat. However, they also highlight opportunities for improvement, such as better capturing fine-grained details and enhancing object recognition in complex cooking scenarios.

3. Additional Results

We also evaluate our methods on VTG-LLM [3] to demonstrate its general effectiveness. Differently, for the clip sequence data, we directly leverage video data from ShareGPT4Video [2]. Following the same training strategy, we first train VTG-LLM using our Seq2TIme with 1 epoch and the rest training follows the original setting of VTG-LLM. As shown in Table 1, Seq2Time brings a 5.23% improvement on the F1 score and 2.37% improvement on the R@0.5 metric, indicating consistent effectiveness for different video LLMs.

	YouCook2				Charades-STA	
	S	**C**	**M**	**F1**	**R@0.5**	**R@0.7**
VTG-LLM*	1.5	**5.1**	1.8	17.2	33.8	15.7
+ S2T	**1.6**	**5.1**	**2.0**	**18.1**	**34.6**	**16.1**

Table 1. The performance of VTG-LLM with Seq2Time as additional training data. *Reproduced VTG-LLM results.

4. Limitation Analysis

From Figure 9 to Figure 14, we observe the overall effectiveness of our Seq2Time in capturing key events and providing accurate temporal annotations. However, some limitations remain, which could be addressed in future work.

First, while the model generates correct captions with precise timestamps, the described events are occasionally less critical within the context of the video. For example, in Figure 12, "0.0 - 40.4 seconds, A pair of girls are preparing a meal in the kitchen," and in Figure 13, "158.7 - 166.4 seconds, The woman takes a bite of the

cake and smiles at the camera," both describe valid events, but these are not essential steps in the respective cooking procedures.

Second, the model sometimes overlooks fine-grained steps that are crucial for understanding detailed processes. For instance, in Figure 11, the model predicts a general event: "33.6 - 103.1 seconds, The man prepares the ingredients for the dish and stirs them together in a pan." While this is accurate at a high level, it misses finer details such as "put olive oil," "drain off the fat," and "add chopped vegetables," which are key preparatory steps in making tacos. Addressing these limitations would enhance the granularity and relevance of the captions generated by Seq2Time.

In the image, there are two individuals standing in front of a white self-driving car. The car is equipped with a roof-mounted sensor array, which is characteristic of autonomous vehicles designed to navigate roads without human intervention. The license plate of the car reads "6WRE502". The person on the left is wearing a blue t-shirt and sunglasses, while the person on the right is dressed in a black blazer and a white shirt. Both individuals are smiling and appear to be posing for the photo. In the background, there are trees and a building, suggesting that the photo was taken in an urban or suburban setting. There is also a motorcycle parked to the left of the car. The overall scene suggests a casual, outdoor setting, possibly during a demonstration or event related to self-driving technology.

Figure 1. Illustration of an example from LLaVA-ReCap-CC3M. The caption provide details in every aspects of the image from foreground to background. For instance, even the motorcycle in the background is captured. The important descriptions are indicated in green texts.

Question Templates for Image Index Grounding with Single Target:

"Which image matches the description: {}? Please output the image index."
"Identify the image that corresponds to this caption: {}. Please provide the index."
"What is the index of the image that describes {}?"
"Can you find the image that matches this description: {}? Please give the index."
"Which image is described by {}? Please output the index."
"Find the image that matches the following description: {}. Please output the index."
"What is the index of the image that corresponds to {}?"
"Which image aligns with the caption: {}? Please provide the index."
"Identify the image matching this description: {}. Please give the index."
"What image index corresponds to the description: {}?"

Answer Templates for Image Index Grounding with Single Target:

"The image index is {}."
"You should look at image number {}."
"The correct index is {}."
"Check image of index {} for the answer."
"It corresponds to image of index {}."
"Image number {} is the correct one."
"The image matching this description is at index {}."
"You will find it at image {}."
"The right image is at index {}."
"Refer to image {}."

Figure 2. Instruction templates of image index grounding with single target image.

Question Templates for Image Index Grounding with Multiple Targets:

"Which images match the following descriptions: ",
"Identify the images corresponding to these captions: ",
"What are the indices of the images that describe the following: ",
"Can you find the images that match these descriptions: ",
"Which images are described by the following: ",
"Find the images matching these descriptions: ",
"What are the image indices that correspond to these captions: ",
"Which images align with the captions: ",
"Identify the images matching these descriptions: ",
"What image indices correspond to these descriptions: "

Following Prompt after Question:

The 1st caption is {}; the 2nd caption is {}; … **<omitted>**
Please output the image indices in the same order as the captions.

Answer Templates for Image Index Grounding with Multiple Targets:

"The image indices are {}.",
"You should look at image numbers {}.",
"The correct indices are {}.",
"Check images with indices {} for the answer.",
"They correspond to images of indices {}.",
"Image numbers {} are the correct ones.",
"The images matching these descriptions are at indices {}.",
"You will find them at images {}.",
"The right images are at indices {}.",
"Refer to images of indices {}."

Figure 3. Instruction templates of image index grounding with multiple target images.

Question Templates for Indexed Image Captioning with Single Target:

"Please describe the image with index {}."
"What does the image with index {} depict?"
"Can you provide the description for the image at index {}?"
"Describe the image that is indexed as {}."
"What is the caption for the image with index {}?"
"Tell me what the image at index {} shows."
"Could you describe the content of the image at index {}?"
"What does the image numbered {} describe?"
"Provide the description for the image at index {}."
"What is depicted in the image with index {}?"

Answer Templates for Indexed Image Captioning with Single Target:

"The image with index {} describes {}."
"At index {}, the image depicts {}."
"The image at index {} shows {}."
"Index {} corresponds to an image that describes {}."
"The description for the image at index {} is {}."
"For the image indexed at {}, the caption is {}."
"The image numbered {} describes {}."
"At index {}, the image is about {}."
"The caption for the image at index {} is {}."
"The image at index {} depicts {}."

Figure 4. Instruction templates of indexed image captioning with single target image.

Question Templates for Indexed Image Captioning with Multiple Targets :

"Please describe the images with the following indices: "
"What do the images at the following indices depict: "
"Can you provide descriptions for the images indexed as: "
"Describe the images corresponding to these indices: "
"What are the captions for the images with indices: "
"Tell me what the images at the following indices show: "
"Could you describe the content of the images at these indices: "
"Provide the descriptions for the images at these indices: "
"What is depicted in the images with the following indices: "

Answer Templates for Indexed Image Captioning with Multiple Targets:

"The image with index {} describes: {}."
"At index {}, the image depicts: {}."
"The image at index {} shows: {}."
"Index {} corresponds to an image that describes: {}."
"The description for the image at index {} is: {}."
"For the image indexed at {}, the caption is: {}."
"The image numbered {} describes: {}."
"At index {}, the image is about: {}."
"The caption for the image at index {} is: {}."
"The image at index {} depicts: {}."

Figure 5. Instruction templates of indexed image captioning with multiple target images.

Question Templates for Adjacent Location Reasoning:

"Which image comes right before the one depicting {}? Please output the index and describe this image."
"Which image comes right after the one depicting {}? Please output the index and describe this image."
"What is the image right before the one showing {}? Please output the index and describe the target image."
"What is the image right after the one showing {}? Please output the index and describe the target image."
"Which image is right before the one describing {}? Please output the index and describe this image."
"Which image is right after the one describing {}? Please output the index and describe this image."
"What is the image right before the one showing {}? Please provide the index and describe the image."
"What is the image right after the one showing {}? Please provide the index and describe the image."
"Identify the image that comes immediately before the one featuring {}. Please output the index and describe it."
"Identify the image that comes immediately after the one featuring {}. Please output the index and describe it."
"Which image is right before the one depicting {}? Please provide the index and describe the image."
"Which image is right after the one depicting {}? Please provide the index and describe the image."
"What is the image immediately before the one showing {}? Please output the index and describe this image."
"What is the image immediately after the one showing {}? Please output the index and describe this image."
"Which image is right before the one describing {}? Please output the index and describe the target image."
"Which image is right after the one describing {}? Please output the index and describe the target image."
"What is the image right before the one showing {}? Please provide the index and describe it."
"What is the image right after the one showing {}? Please provide the index and describe it."
"Identify the image that comes right before the one featuring {}. Please output the index and describe the image."
"Identify the image that comes right after the one featuring {}. Please output the index and describe the image."

Answer Templates for Adjacent Location Reasoning:

"Refer to image {}. It shows {}.",
"Look at image {}. It depicts {}.",
"Image {} describes {}.",
"You should check image {}. It illustrates {}.",
"The correct image is {}. It shows {}.",
"Image {} presents {}.",
"The target image is {}. It depicts {}.",
"Refer to image {}. It describes {}.",
"You should see image {}. It illustrates {}.",
"Image {} shows {}.",

Figure 6. Instruction templates of adjacent location reasoning.

The video captures an intense moment during a basketball game. A player in an orange jersey is leaping towards the hoop, attempting to dunk or block the ball with a powerful leap, while other players from both teams are positioned strategically around the court, ready for the play's outcome. The crowd in the stands is visible, their faces a blur of anticipation and excitement as they watch the action unfold. The lighting inside the arena highlights the drama of the game, casting dramatic shadows and emphasizing the athleticism of the players on the court.

Figure 7. Instruction of an example in K700-LongVA-Captions.

In this video, ***a group of adults and babies*** *are gathered around a table, engrossed in reading a colorful newspaper. The adults and children are* ***all wearing green*** *hats with a distinctive leaf design on them, suggesting a theme or celebration related to spring or nature. The atmosphere is warm and familial, with* ***each person holding a baby*** *securely as they share the experience of reading together. The room is filled with homely warmth, and the focus is clearly on the shared joy of learning and spending time together.*

Figure 8. Instruction of a low-quality example in K700-LongVA-Captions. The incorrect descriptions are indicated by **bold red** text.

References

[1] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019. 1

[2] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, et al. Sharegpt4video: Improving video understanding and generation with better captions. *arXiv preprint arXiv:2406.04325*, 2024. 2

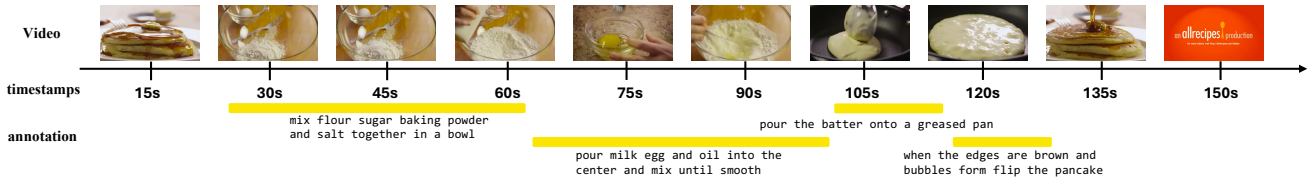
[3] Yongxin Guo, Jingyu Liu, Mingda Li, Xiaoying Tang, Xi Chen, and Bo Zhao. Vtg-llm: Integrating timestamp knowledge into video llms for enhanced video temporal grounding. *arXiv preprint arXiv:2405.13382*, 2024. 2

[4] Bo Li, Hao Zhang, Kaichen Zhang, Dong Guo, Yuanhan Zhang, Renrui Zhang, Feng Li, Ziwei Liu, and Chunyuan Li. Llava-next: What else influences visual instruction tuning beyond data?, 2024. 1

[5] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024. 1

[6] OpenAI. Gpt-4 technical report. https://openai.com/research/gpt-4, 2023. Accessed: 2024-11-07. 1

[7] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024. 1



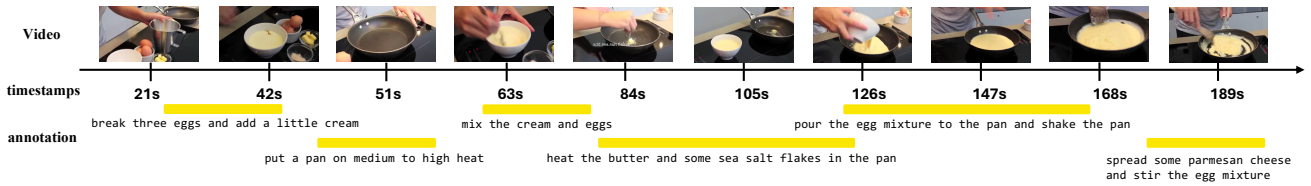
TimeChat

0.0 - 21.3 seconds, We see the words blueberry ice cream and the words ingredients.
21.3 - 43.5 seconds, We see a bowl of eggs and the words vanilla and salt.
43.5 - 60.5 seconds, We see a bowl of powdered sugar and the words powdered sugar and blueberries.
60.5 - 85.3 seconds, We see a pan with melted butter and blueberries in a bowl.
85.3 - 104.3 seconds, We see a bowl of flour and the words baking powder.
104.3 - 122.5 seconds, We see a pan with a blueberry muffin and we are shown the ingredients.
122.5 - 143.3 seconds, We see a bowl of powdered sugar and blueberries.
143.3 - 148.8 seconds, We see the finished blueberry muffin.

+Seq2Time

0.0 - 16.3 seconds, The video starts by showing a plate with a variety of blue pancakes on it.
16.3 - 25.0 seconds, The next step is to prepare the ingredients needed to make the blue pancakes.
25.0 - 61.4 seconds, The next step is to mix the ingredients together in a bowl.
61.4 - 89.2 seconds, The next step is to pour the mixture into a pan and cook it on a stove.
89.2 - 120.7 seconds, The next step is to let the pancakes cook until they are done.
120.7 - 141.5 seconds, The next step is to take the pancakes out of the pan and place them on a plate.
141.5 - 149.0 seconds, The video ends with the words \"Subscribe to our channel\".

Figure 9. Qualitative example of our Seq2Time on TimeChat. The incorrect descriptions and timestamps are indicated by brown text.



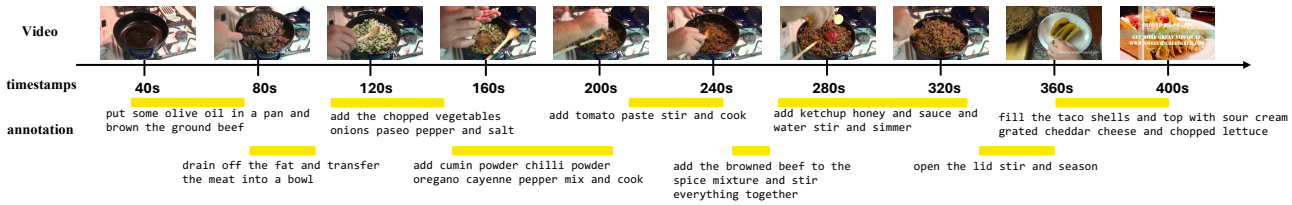
TimeChat

0.0 - 11.8 seconds, A blue woman is standing at a stove talking.
11.8 - 17.3 seconds, She has a blue colored bowl in her hands.
17.3 - 19.3 seconds, She sets the bowl on the stove.
19.3 - 40.8 seconds, She picks up a container and pours something into the bowl.
40.8 - 46.7 seconds, She begins stirring the mixture with a spoon.
46.7 - 60.8 seconds, She pours more into the bowl.
60.8 - 84.6 seconds, She begins to mix the mixture more.
84.6 - 113.6 seconds, She pours more into the bowl and mixes it in with a spoon.
113.6 - 124.2 seconds, She takes a spoon and stirs it.
124.2 - 142.2 seconds, She pours some into the bowl.
142.2 - 153.6 seconds, She pours more into the bowl and stirs it.
153.6 - 158.9 seconds, She pours some into the bowl and stirs it.
158.9 - 166.3 seconds, She takes a spoon and stirs it.
166.3 - 173.2 seconds, She pours some into the bowl and stirs it.
173.2 - 181.2 seconds, She pours some into the bowl and stirs it.
181.2 - 193.1 seconds, She pours some into the bowl and stirs it.
193.1 - 197.6 seconds, She pours some into the bowl and stirs it.
197.6 - 204.0 seconds, She pours some into the bowl and stirs it.
204.0 - 208.1 seconds, She pours some into the bowl and stirs it.
208.1 - 211.3 seconds, She pours some into the bowl and stirs it.

+Seq2Time

0.0 - 9.4 seconds, A woman is seen standing in a kitchen with a bowl of eggs in front of her.
29.4 - 54.6 seconds, She proceeds to crack the eggs into a pan with salt and water.
54.6 - 74.4 seconds, She then proceeds to stir the mixture in the pan.
74.4 - 84.5 seconds, She stirs the mixture again.
84.5 - 113.8 seconds, She adds more salt to the mixture in the pan.
113.8 - 154.6 seconds, She then proceeds to cook the mixture in the pan and stir it.
154.6 - 170.3 seconds, She adds a pancake mix to the pan and proceeds to stir it in the pan.
170.3 - 187.8 seconds, She continues to stir the mixture until it is finished.

Figure 10. Qualitative example of our Seq2Time on TimeChat. The incorrect descriptions and timestamps are indicated by brown text and repetitive text is highlighted in red.



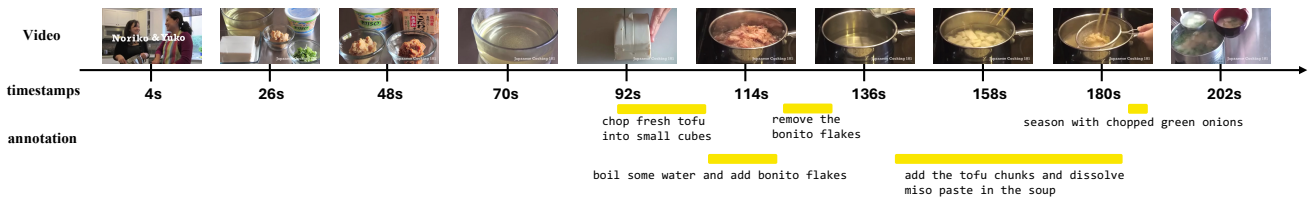
TimeChat

0.0 - 17.9 seconds, A man in a kitchen wearing a blue shirt with white background shows a bowl and a blue cloth.
17.9 - 39.3 seconds, He pours some blue powder into the bowl and mix it with water to create a solution.
39.3 - 406.1 seconds, He pours the solution on a dish, covers it with a plastic sheet and leaves it to settle.
406.1 - 414.3 seconds, He then prepares another dish with different ingredients and mix them in a bowl.
414.3 - 426.5 seconds, He pours the mixture on a dish and places it on the table.
426.5 - 453.6 seconds, He then takes the dish and spreads a white powder on it and places it on a baking tray.
453.6 - 466.6 seconds, He spreads some white powder on another dish and places it in the oven to bake.
466.6 - 482.3 seconds, He takes a blue cloth and wraps it on the dish and leaves it in the oven to bake.
482.3 - 494.9 seconds, He then takes the dish out of the oven and wraps it in a plastic sheet.
494.9 - 498.2 seconds, He takes the dish out of the oven and unwraps it.
498.2 - 506.0 seconds, He then places the dish on a table and unwraps the plastic sheet.
506.0 - 512.6 seconds, He then adds some sauce on the dish and serves it.

+Seq2Time

0.0 - 33.6 seconds, The man introduces himself and begins talking about the dish he will be making.
33.6 - 103.1 seconds, The man prepares the ingredients for the dish and stirs them together in a pan.
103.1 - 148.1 seconds, The man cooks the dish for a few minutes and then stirs in more ingredients.
148.1 - 259.2 seconds, The man adds more ingredients to the dish and continues to stir it together.
259.2 - 300.3 seconds, The man finishes cooking the dish and begins to serve it.

Figure 11. Qualitative example of our Seq2Time on TimeChat. The incorrect descriptions and timestamps are indicated by brown text.



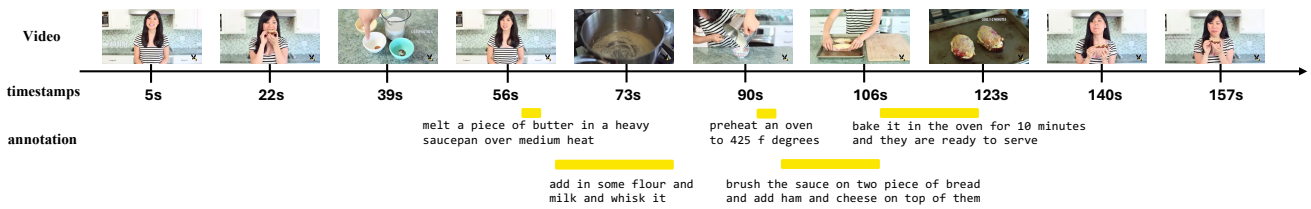
TimeChat

0.0 - 19.4 seconds, Two girls are standing in front of a table and they're talking to each other.
19.4 - 23.0 seconds, Then, a man starts to prepare the ingredients for cooking a dish.
23.0 - 227.8 seconds, The man is kneeling in front of a bowl of water and he's putting salt on it.
227.8 - 231.4 seconds, After, the man put salt in the water, then he takes the salt out of the water.
231.4 - 237.4 seconds, The man is kneeling in front of a pan and he's putting the salt on the pan.
237.4 - 242.2 seconds, After, the man takes the salt out of the pan.

+Seq2Time

0.0 - 40.4 seconds, A pair of girls are preparing a meal in the kitchen.
40.4 - 70.8 seconds, They mix ingredients in a bowl.
70.8 - 136.1 seconds, The girls put the bowl in the microwave and melt ice.
136.1 - 211.4 seconds, They mix the ingredients in a pot and stir it.

Figure 12. Qualitative example of our Seq2Time on TimeChat. The incorrect descriptions and timestamps are indicated by brown text.



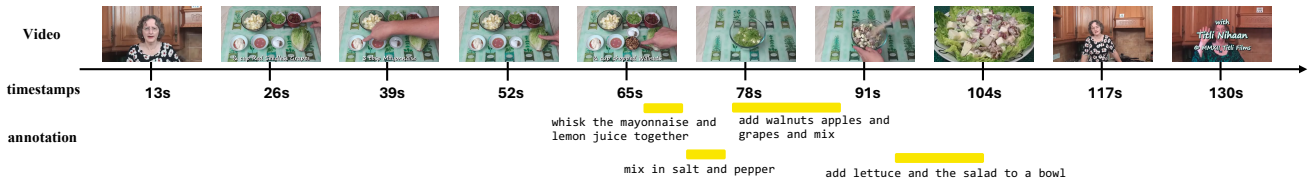
TimeChat

0.0 - 16.4 seconds, A young man is sitting in front of a camera.
16.4 - 24.3 seconds, He is holding a bowl in his hand.
24.3 - 41.4 seconds, He pours blue colored substance into the bowl.
41.4 - 48.2 seconds, He is kneading the substance with his hands.
48.2 - 60.3 seconds, He takes the mixture out of the bowl and spreads it over a baking tray.
60.3 - 75.6 seconds, He then places another baking tray on top of the baking tray.
75.6 - 92.4 seconds, He places a toothbrush in the mixture and then puts a plate on top.
92.4 - 114.6 seconds, He is then pouring blue liquid into the baking tray.
114.6 - 134.6 seconds, He then takes out a plate with blue liquid on it and starts eating it.
134.6 - 159.1 seconds, He then places a toothbrush into the blue liquid.
159.1 - 164.3 seconds, He takes out a plate with blue liquid on it and starts eating it.

+Seq2Time

0.0 - 13.0 seconds, A woman wearing blue makeup is seen talking to the camera in a kitchen.
14.0 - 56.0 seconds, The woman is seen preparing a cake in the kitchen by mixing ingredients in bowls and pouring them into a pan.
56.0 - 157.7 seconds, The woman is seen pouring melted margarine onto two slices of bread, then adding toppings onto the bread.
158.7 - 166.4 seconds, The woman takes a bite of the cake and smiles at the camera.

Figure 13. Qualitative example of our Seq2Time on TimeChat. The incorrect descriptions and timestamps are indicated by brown text.



TimeChat

0.0 - 27.0 seconds, A woman stands in a kitchen and discusses salads.
27.0 - 51.7 seconds, A plate of ingredients are shown.
51.7 - 105.3 seconds, A woman fills a cabinet with containers.
105.3 - 131.2 seconds, She takes a container out and begins making a salad.

+Seq2Time

0.0 - 36.4 seconds, A woman talks about lunch while she shows a recipe.
36.4 - 131.1 seconds, She is shown the different ingredients needed and her hands are shown dipping into the ingredients.

Figure 14. Qualitative example of our Seq2Time on TimeChat. The incorrect descriptions and timestamps are indicated by brown text.