ShowHowTo: Generating Scene-Conditioned Step-by-Step Visual Instructions

Supplementary Material

Overview

In the supplementary material, we first provide dataset collection details and show examples from our dataset in Section A. Then, in Section B, we compare our dataset to other related works and discuss their differences. We provide implementation details in Section C and evaluation details and analysis of our metrics in Section D. In Section E, additional quantitative results are provided, and in Section F, we show a large variety of qualitative results.

A. Dataset Collection Details

Speech transcription: YouTube ASR vs. WhisperX. The original video transcripts released with the HowTo100M dataset, generated through YouTube ASR, are known to contain errors [6, 10]. Following Li *et al*. [10], we employ WhisperX [2] to obtain higher-quality transcripts for narrated instructional videos. As shown in Figure 12, WhisperX provides improved transcription quality over YouTube ASR through better punctuation, fewer transcription errors (*e.g*., *"Put some aluminum foil in there"* vs. YouTube ASR's incorrect *"it's a moment of oil in there"*), and more accurate sentence segmentation with timestamps. These improvements are essential for our subsequent dataset creation steps, which rely on accurate transcription.

Details on filtering of irrelevant videos. To address the presence of non-instructional content in HowTo100M, we prompt Llama 3 [4] (`Llama-3.1-8B-Instruct`) to identify whether videos are instructional or not based on their transcript excerpts. The full prompt used is shown in Figure 13.

We validate our approach for video filtering by manually annotating a balanced validation set of 200 videos. Our approach achieves the false positive rate of 5% and the false negative rate of 12%, *i.e*., 95% of non-instructional videos spuriously present in the HowTo100M dataset are filtered out by our automatic approach. Applied to the full HowTo100M dataset, we identified 847K (68.4%) instructional videos and filtered out 391K (31.6%) non-instructional content, including product reviews (*e.g*., *"Houseplant Unboxing — Steve's Leaves"*), entertainment videos (*e.g*., *"Don McLean - American Pie (with Lyrics)"*), and personal vlogs (*e.g*., *"Driving to the West, an RV lifestyle vlog"*).

Step extraction details. Our approach uses Llama 3 to extract the steps from video narrations. This contrasts the related work [18], which extracts steps from video frames and uses an image-captioning model to generate step captions.

Figure 1. **Frame matching comparison across CLIP, SigLIP, and DFN-CLIP (used in our work).** For each method, the figure shows the best matching frame to the instructional text (shown below).

We found that the latter approach resulted in captions that were too brief, high-level, and lacked sufficient detail to differentiate adjacent steps.

Figure 14 illustrates our Llama 3 prompt for extracting instructional steps from video narrations, including the few-shot examples used (truncated to fit into a page; complete prompts will be released with our code). We processed videos under 10 minutes long, as longer transcripts exceeded the context limit of Llama 3 and also degraded output quality. The model is instructed to generate temporally ordered steps with approximate timestamps using video transcripts. We filter out malformed results with non-temporally ordered steps or incomplete descriptions. Examples of extracted steps are shown in Figure 2.

Cross-modal frame alignment details. For each extracted step, we find its matching frame in the video using DFN-CLIP [5] (`DFN5B-CLIP-ViT-H-14-378`), restricting the search to frames within the step's time bounds generated by Llama 3. We formulate this as a dynamic programming

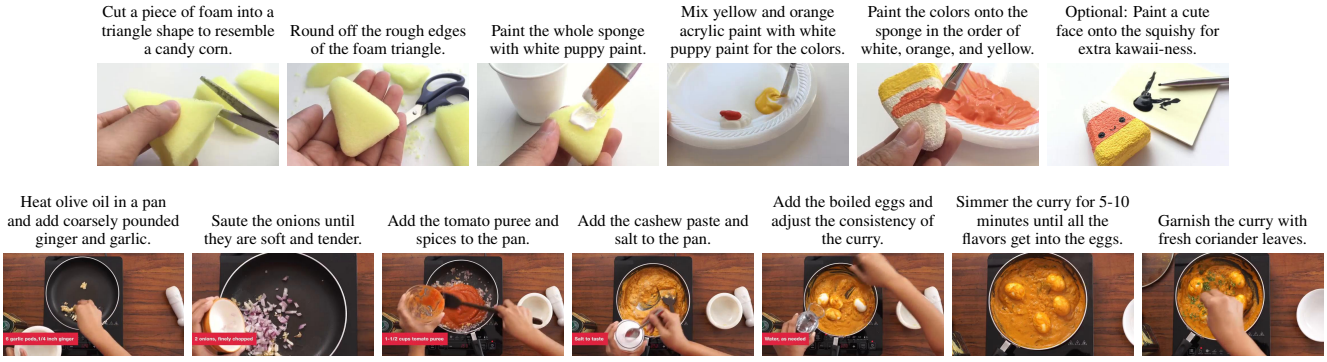

Figure 2. **Samples from the ShowHowTo dataset**. Each sample (row) is a sequence of textual instructions (top) and the associated visual instruction images (bottom).

18.05 - 19.53 ✓ Put aluminum foil in the pressure cooker.
28.86 - 29.70 ✓ Create two little rings using the aluminum foil.
38.50 - 40.42 ✓ Place the potatoes in the pressure cooker, above the water.
44.99 - 50.23 ✓ Add 1.5 cups of water to the pressure cooker.
59.85 - 62.05 ✓ Cook the potatoes for 15 minutes, depending on their size.
76.08 - 80.48 ? Remove the potatoes from the pressure cooker and serve.

4.00 - 12.00 ? Put potatoes in the pressure cooker and create steaming rings.
17.00 - 25.00 ? Add one and a half cups of water to the pressure cooker.
19.00 - 27.00 ✓ Place aluminum foil in the pressure cooker.
78.00 - 86.00 ✗ It doesn't heat up your kitchen.
83.00 - 91.00 ✗ It's really good for in the summer when you want to have something like a baked potato.
86.00 - 94.00 ✗ Up anything else.
91.00 - 99.00 ✗ When they're done, you just pop
96.00 - 104.00 ✗ It's just going to be steamed.
96.00 - 104.00 ✗ So there it is.

✓ – *an instruction with correct approximate timestamp* ? – *an instruction with incorrect timestamp* ✗ – *not an instruction*

Figure 3. **Comparison between textual instructions** extracted by our method (left) and the textual instructions from HowToStep [10] (right) for the same randomly chosen *'How to bake a potato in the pressure cooker'* video. The original transcript used to produce our instructions is shown in Figure 12. Our method correctly identifies the key steps in the narrations and summarizes them in step-by-step instructions. On the other hand, the HowToStep data often contain steps that are not instructions.

problem to find optimal frame-text pairs while preserving temporal order and maximizing alignment scores.

Although the temporal boundaries generated by Llama 3 are rarely incorrect, narrations do not always align with respect to video frames and can occur slightly before/after the visuals. To account for the well-known temporal misalignment [6, 10], we expand the temporal boundaries by a fixed duration of ϵ seconds, increasing the search space for frame alignment. Through analysis on a small validation set of manually annotated videos with precise step boundaries, we found $\epsilon = 15$ seconds yielded a good balance of precision (the correct frame was selected by DFN-CLIP) and recall (the correct frame was inside the searched interval).

We use DFN-CLIP over related contrastive models such as CLIP [14] and SigLIP [22] as we found it superior in certain cases for matching video frames to instructional steps. We observed CLIP and SigLIP often exhibited limitations such as incorrect object state identification (*e.g.*, unroasted peppers, uncut prosciutto in Figure 1, rows 1-2) and tendency to select blurry or transition frames (Figure 1, rows 2-3). We quantify the quality of the automatically selected keyframes by a small user study. We manually annotated keyframes for 100 steps and asked humans to blindly select whether the manually or automatically selected frame better corresponds to the text instruction. DFN-CLIP was preferred in 18% of cases, human annotation was preferred in 36% of cases, remaining 46% of cases were a tie. This indicates fairly decent alignment with human judgment.

Dataset Statistics. We analyze the ShowHowTo dataset statistics in Figures 4 and 5. The dataset encompasses 25K tasks across diverse categories, including Food and Entertainment, Hobbies and Crafts, *etc*., derived from HowTo100M's task hierarchy. Figure 4 (left) shows the distribution of these categories in our dataset. Each sample contains an average of 7.7 steps, with 11.37 words per step (Figure 4, right). The word clouds in Figure 5 showcases common verbs of physical actions like *remove*, *add*, and *make*, alongside various household objects and materials used in everyday tasks.

B. Relation to Existing Datasets

We show the comparison of related instructional datasets in Table 1. Early datasets like CrossTask [23] and COIN [19] are manually curated but small in scale with categorical instruction annotations. While recent datasets like HowTo-Caption [16] have expanded significantly (1.1M sequences), they provide generic captions rather than instructional text annotations. Specialized datasets exist for egocentric domains (LEGO [9], Ego4D Goal-Step [17]) and single-step

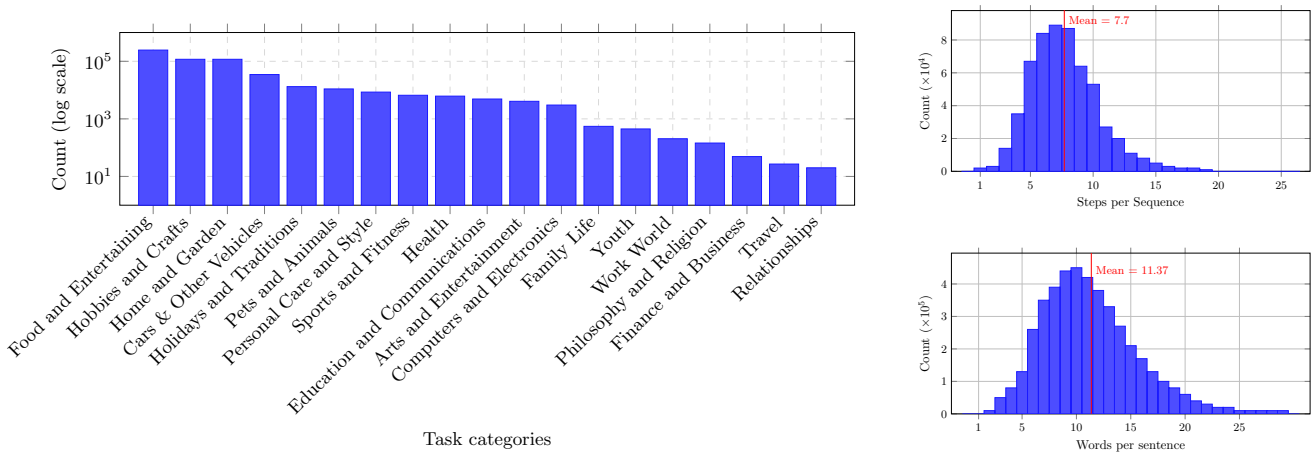

Figure 4. **Statistics of the ShowHowTo dataset**. Left: Distribution of task categories in ShowHowTo dataset. Top-right: Distribution of the number of steps per sequence. Bottom-right: Distribution of the number of words per sentence.

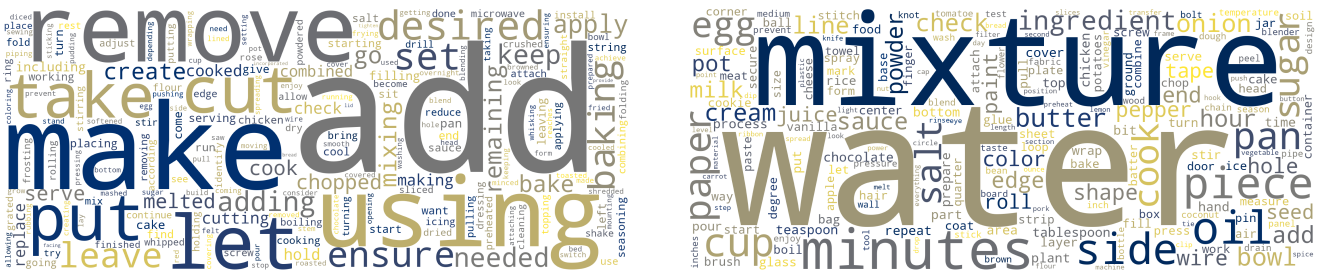

Figure 5. **Word cloud visualizations** of verbs (left) and nouns (right) in the textual instructions of ShowHowTo dataset.

instructions (AURORA [8], GenHowTo [18]), but their narrow scope limits general applicability.

The two most comparable large-scale multi-step instructional datasets to our dataset are WikiHow-VGSI [21] (100K sequences) and HowToStep [10] (312K sequences). WikiHow-VGSI, composed of image-step pairs extracted from WikiHow articles, predominantly contains digitally drawn illustrations rather than real photos, making it unsuitable for realistic image generation and very difficult to scale. On the other hand, HowToStep [10], similarly to us, leverages HowTo100M videos, but its focus is solely on the cooking domain, and it contains only temporal video segments instead of individual representative images. Lastly, the provided segments are of variable quality, often not being instructional (see Figure 3 for an example). In comparison, the ShowHowTo dataset is both larger in scale and contains more diverse tasks with higher-quality steps. In the main paper, we train our model on both WikiHow-VGSI and HowToStep, showing that training on our dataset results in models with substantially better generation capabilities.

C. Implementation Details

ShowHowTo implementation details. We initialize our model with pretrained weights of DynamiCrafter for image animation [20] and train the model for approximately 100K steps on four AMD MI250x multi-chip modules (8 GPUs) with a total of 512 GB of VRAM (64 GB per GPU). We use the batch size of 16 image sequences at the resolution of 256×256 pixels. The image sequence length is variable, ranging between two and eight. We train the model using AdamW optimizer with a learning rate of $2{\cdot}10^{-5}$. The training takes approximately 48 hours. During inference, we generate visual instruction sequences using the DDIM sampler with 50 denoising steps. Both the inference and training code, along with trained model weights will be made publicly available.

Implementation details of the related methods. In our evaluation of the related methods, we use the official implementations of GenHowTo, AURORA, and InstructPix2Pix provided by the respective authors. In the case of GenHowTo, we use the action model weights. For StackedDiffusion [11], we use the official implementation modified to replace the first generated image with the noised version of

Dataset	Source	Manually Curated	Task Domain	Visual Domain	Scale	# Tasks	Avg. Steps / Seq.	Visual Type	Annotation Type
CrossTask [23]	YouTube	✓	Open	Open	4.7K	18	7.4	Video Segments	Categorical
COIN [19]	YouTube	✓	Open	Open	10K	180	3.9	Video Segments	Categorical
Ego4D Goal-Step [17]	Ego4D	✓	Open	Egocentric	717‡	80	23.3	Video Segments	Instructions
LEGO [9]	Ego4D & EPIC-K.	✗	Open	Egocentric	147K	-	1.0	Key Frames	Instructions
AURORA [8]	Multiple	✓	Open	Open	289K	-	1.0	Key Frames	Instructions
GenHowTo [18]	COIN & ChangeIt	✗	Open	Open	45K	224	2.0	Key Frames	Captions
HowToCaption [16]	HowTo100M	✗	Open	Open	1.1M	23.6K	18.5	Video Segments	Captions
HT-Step [1]	HowTo100M	✓	Cooking	Open	18K	433	5.9	Video Segments	Instructions
HowToStep [10]	HowTo100M	✗	Cooking	Open	312K	14.2K	10.6	Video Segments	Instructions
WikiHow-VGSI [21]	WikiHow	✓	Open	Illustrations†	100K	53.2K	6.0	Key Frames	Instructions
ShowHowTo	HowTo100M	✗	Open	Open	578K	25K	7.7	Key Frames	Instructions

† Some examples in the dataset are real photos. ‡ The subset with step annotations.

Table 1. **Comparison of instructional datasets**. Scale refers to the number of instruction sequences of image-text pairs. Annotation Type describes the nature of text captions. Visual Type indicates the format of visual content.

the input image in each denoising step, to enable input image conditioning. The work of Phung *et al*. [13] does not provide the official implementation at the time of writing; therefore, we reimplement it using the Stable Diffusion 2.1 model. For the step similarity matrix, we use the matrix with all values set to one, and similarly to the StackedDiffusion model, we replace the first generated image with the noised version of the input image in each denoising step.

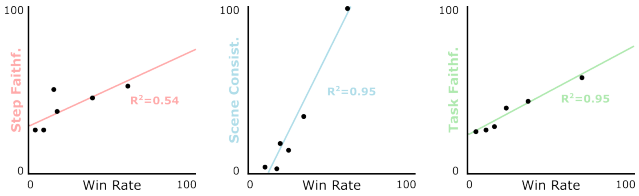

Figure 6. **Correlation of our metrics with human preference.** User study results (Win Rate %) highly correlate with our three metrics—Step Faithfulness, Scene Consistency, and Task Faithfulness.

Method	FID↓
InstructPix2Pix [3]	37.8
AURORA [8]	23.2
GenHowTo [18]	28.3
Phung *et al*. [13]	27.8
StackedDiffusion [11]	34.6
ShowHowTo	**12.4**

Table 2. **Comparison with state-of-the-art using the FID score** on the ShowHowTo test set. ShowHowTo model significantly outperforms all related methods.

Method	Acc_{ac}
Stable Diffusion [15]	0.51
Edit Friendly DDPM [7]	0.60
InstructPix2Pix [3]	0.55
GenHowTo [18]	0.66
ShowHowTo	**0.72**

Table 3. **Zero-shot evaluation on the GenHowTo dataset** according to the GenHowTo protocol [18]. The ShowHowTo model outperforms the prior state-of-the-art with no fine-tuning on the GenHowTo train set.

D. Evaluation Details

As described in the main paper, we evaluate our method using three metrics: Step Faithfulness, Scene Consistency, and Task Faithfulness. DFN-CLIP [5] (`DFN5B-CLIP--ViT-H-14-378`) is used for the computation of the Step Faithfulness and Task Faithfulness metrics. Scene Consistency is computed with the averaged spatial patch features of DINOv2 [12] (`dinov2_vitb14_reg`). All metrics are first averaged per sequence before being averaged across the test set to account for the variable sequence length.

We verify how well the used metrics correlate with human preference as measured by our user study (see the main paper). For each evaluation metric, we plot the performance of all other models against the win rate compared to the ShowHowTo model from the user study. The results can be seen in Figure 6, we also show the line of best fit and R^2 value. We find a high correlation across all metrics, especially so for Scene Consistency and Task Faithfulness, confirming that our metrics serve as a good proxy for human preference.

E. Additional Quantitative Results

FID results. We also evaluate all methods using the FID score in Table 2. For each input image I_0 and the textual instructions $\{\tau_i\}_{i=0}^n$ from the ShowHowTo test set, we generate the sequence $\{\hat{I}\}_{i=1}^n$ of visual instructions. For each method, its FID score is computed between its generated sequences and the source visual instruction sequences $\{I_i\}_{i=1}^n$ from the test set. Our method generates images that better match the distribution of real visual instruction sequences.

Zero-shot evaluation on the GenHowTo dataset. In GenHowTo [18], the authors propose to evaluate generative

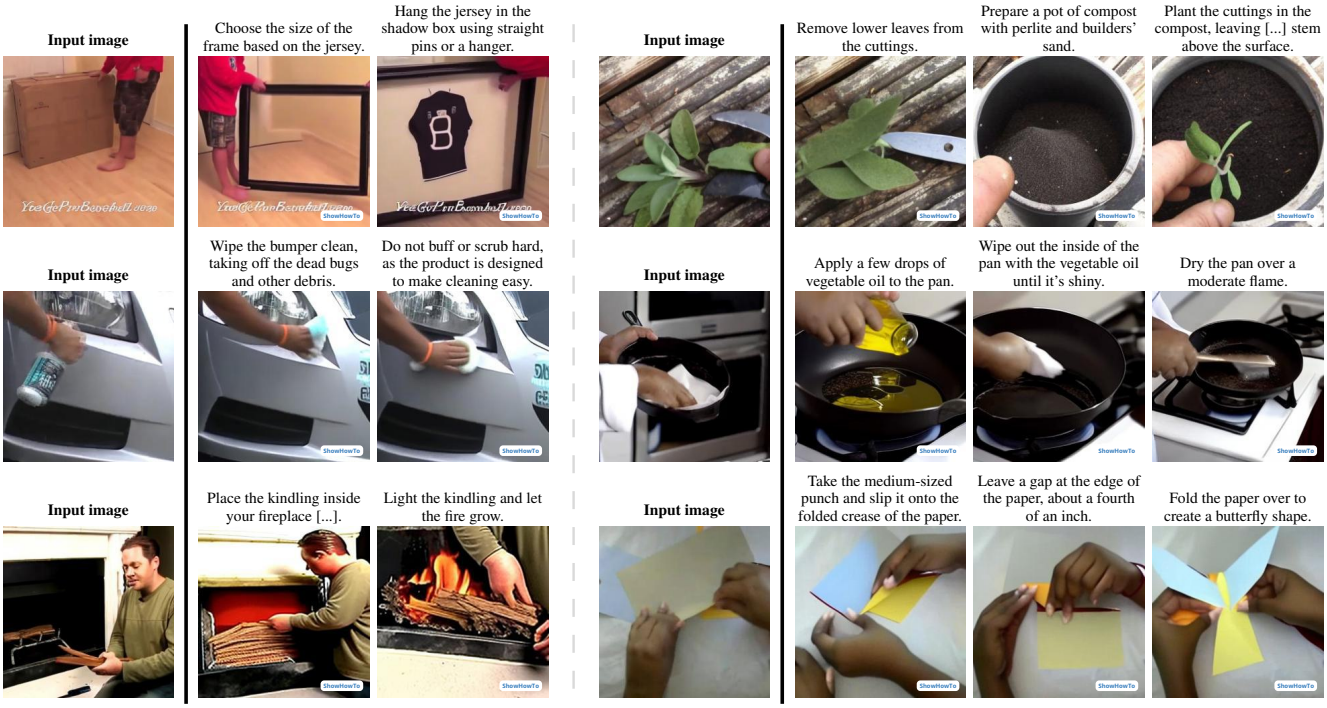

Figure 7. **Qualitative results of our method for sequences from the test set.** Our model can generate both short instructional sequences (as shown here) as well as long or very long sequences shown in Figures 8 and 9.

	Number of generated frames						
	1	2	3	4	5	6	7
Step Faith.	1.00	0.72	0.59	0.52	0.50	0.50	0.51
Scene Consist.	0.20	0.29	0.43	0.42	0.34	0.36	0.31
Task Faith.	0.30	0.49	0.48	0.45	0.42	0.46	0.37

Table 4. **Model's performance for different lengths of the generated sequence.** The performance is fairly similar across different lengths, with a decrease observed for longer sequences.

Task category	Step Faithf.	Scene Consist.	Task Faithf.	Average
Cars & Other Vehicles	0.34	0.37	0.58	0.43
Education and Communications	0.37	0.55	0.24	0.39
Food and Entertaining	0.67	0.24	0.40	0.44
Health	0.29	0.56	0.46	0.44
Hobbies and Crafts	0.39	0.46	0.41	0.42
Holidays and Traditions	0.56	0.35	0.52	0.48
Home and Garden	0.40	0.41	0.38	0.40
Pets and Animals	0.48	0.41	0.39	0.43
Sports and Fitness	0.36	0.42	0.55	0.44

Table 5. **Model's performance for different task categories.** Task categories with only a few sequences not shown.

methods in a downstream application, where the method generates images of various classes that are then used to train a simple classifier. The performance of this classifier is then computed on the real set of images. We show the results of our method in Table 3. We evaluate our model as is, *i.e.*, trained on the ShowHowTo dataset without any additional training or fine-tuning. We report the action accuracy metric Acc_{ac}, which evaluates whether our generated visual instruction images can be used for downstream application of classifying actions. Note that this metric is image-based and does not evaluate sequences. Nonetheless, ShowHowTo improves over the previous state-of-the-art from [18] by 6 percentage points.

Variable sequence length generation analysis. Our model generates sequences of variable length. We analyze the performance of the model for various sequence lengths in Table 4. Except for the degenerative case with one frame, the performance is fairly similar across different lengths, with a decrease observed for longer sequences.

Per-task performance analysis. We report the performance of our model in various HowTo100M task categories in Table 5. We observe that the performance is significantly dependent on the task distribution in the dataset. In detail, for the Step Faithfulness metric, the best performance is achieved in cooking tasks because of plentiful training data and clear, visually distinct steps. On the other hand, for Scene Consistency, the cooking tasks perform the worst as the matching is done across the whole dataset, where a large portion is cooking with many similar scenes and frames. Additionally, the cooking tasks contain many close-up scenes without any scene background that can be used

for matching the generated images to the correct sequence. The best Scene Consistency is achieved for tasks in the Health and the Education categories due to the uniqueness of the sequences in those categories.

F. Additional Qualitative Results

Additional qualitative results. We show additional qualitative results in Figures 7, 8, and 9. We show our method can correctly generate sequences of visual instructions according to the input images and prompts. In Figure 9, we demonstrate our method can generate long instructional sequences while preserving consistency with the input image. Additionally, in Figure 7, we show our model can also generate shorter sequences.

Additional qualitative comparison with related work. We show additional comparison with related work on the task of creating paper flowers in Figure 11. We can see our method not only correctly captures the scene, which is not the case for the method of Phung *et al*. [13] and Stacked-Diffusion [11], but the model faithfully follows the input prompts, generating useful visual instructions for the user.

Failure modes. We show some limitations of our method, as described in the main paper, in Figure 10. Our model can struggle with objects that are not common in the training data, such as engine cylinders and tools such as razor blades. Additionally, the model can make errors in scenarios where object states need to be tracked and updated across multiple frames, such as after cooking meat, the meat's state must be changed from *raw* to *cooked*, *etc*.

References

[1] Triantafyllos Afouras, Effrosyni Mavroudi, Tushar Nagarajan, Huiyu Wang, and Lorenzo Torresani. Ht-step: Aligning instructional articles with how-to videos. *NeurIPS*, 2024. 4

[2] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, 2021. 1, 10

[3] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023. 4

[4] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 1

[5] Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks. *arXiv preprint arXiv:2309.17425*, 2023. 1, 4

[6] Tengda Han, Weidi Xie, and Andrew Zisserman. Temporal alignment networks for long-term video. In *CVPR*, 2022. 1, 2

[7] Inbar Huberman-Spiegelglas, Vladimir Kulikov, and Tomer Michaeli. An edit friendly ddpm noise space: Inversion and manipulations. In *CVPR*, 2024. 4

[8] Benno Krojer, Dheeraj Vattikonda, Luis Lara, Varun Jampani, Eva Portelance, Christopher Pal, and Siva Reddy. Learning action and reasoning-centric image editing from videos and simulations. *arXiv preprint arXiv:2407.03471*, 2024. 3, 4

[9] Bolin Lai, Xiaoliang Dai, Lawrence Chen, Guan Pang, James M Rehg, and Miao Liu. Lego: Learning egocentric action frame generation via visual instruction tuning. *arXiv preprint arXiv:2312.03849*, 2023. 2, 4

[10] Zeqian Li, Qirui Chen, Tengda Han, Ya Zhang, Yanfeng Wang, and Weidi Xie. Multi-sentence grounding for long-term instructional video. 2024. 1, 2, 3, 4

[11] Sachit Menon, Ishan Misra, and Rohit Girdhar. Generating illustrated instructions. In *CVPR*, 2024. 3, 4, 6

[12] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *TMLR*, 2024. 4

[13] Quynh Phung, Songwei Ge, and Jia-Bin Huang. Coherent zero-shot visual instruction generation. *arXiv preprint arXiv:2406.04337*, 2024. 4, 6

[14] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2

[15] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 4

[16] Nina Shvetsova, Anna Kukleva, Xudong Hong, Christian Rupprecht, Bernt Schiele, and Hilde Kuehne. Howtocaption: Prompting llms to transform video annotations at scale. In *ECCV*, 2024. 2, 4

[17] Yale Song, Eugene Byrne, Tushar Nagarajan, Huiyu Wang, Miguel Martin, and Lorenzo Torresani. Ego4d goal-step: Toward hierarchical understanding of procedural activities. *NeurIPS*, 2024. 2, 4

[18] Tomáš Souček, Dima Damen, Michael Wray, Ivan Laptev, and Josef Sivic. Genhowto: Learning to generate actions and state transformations from instructional videos. In *CVPR*, 2024. 1, 3, 4, 5

[19] Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. Coin: A large-scale dataset for comprehensive instructional video analysis. In *CVPR*, 2019. 2, 4

[20] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *ECCV*, 2024. 3

[21] Yue Yang, Artemis Panagopoulou, Qing Lyu, Li Zhang, Mark Yatskar, and Chris Callison-Burch. Visual goal-step inference using wikihow. In *EMNLP*, 2021. 3, 4

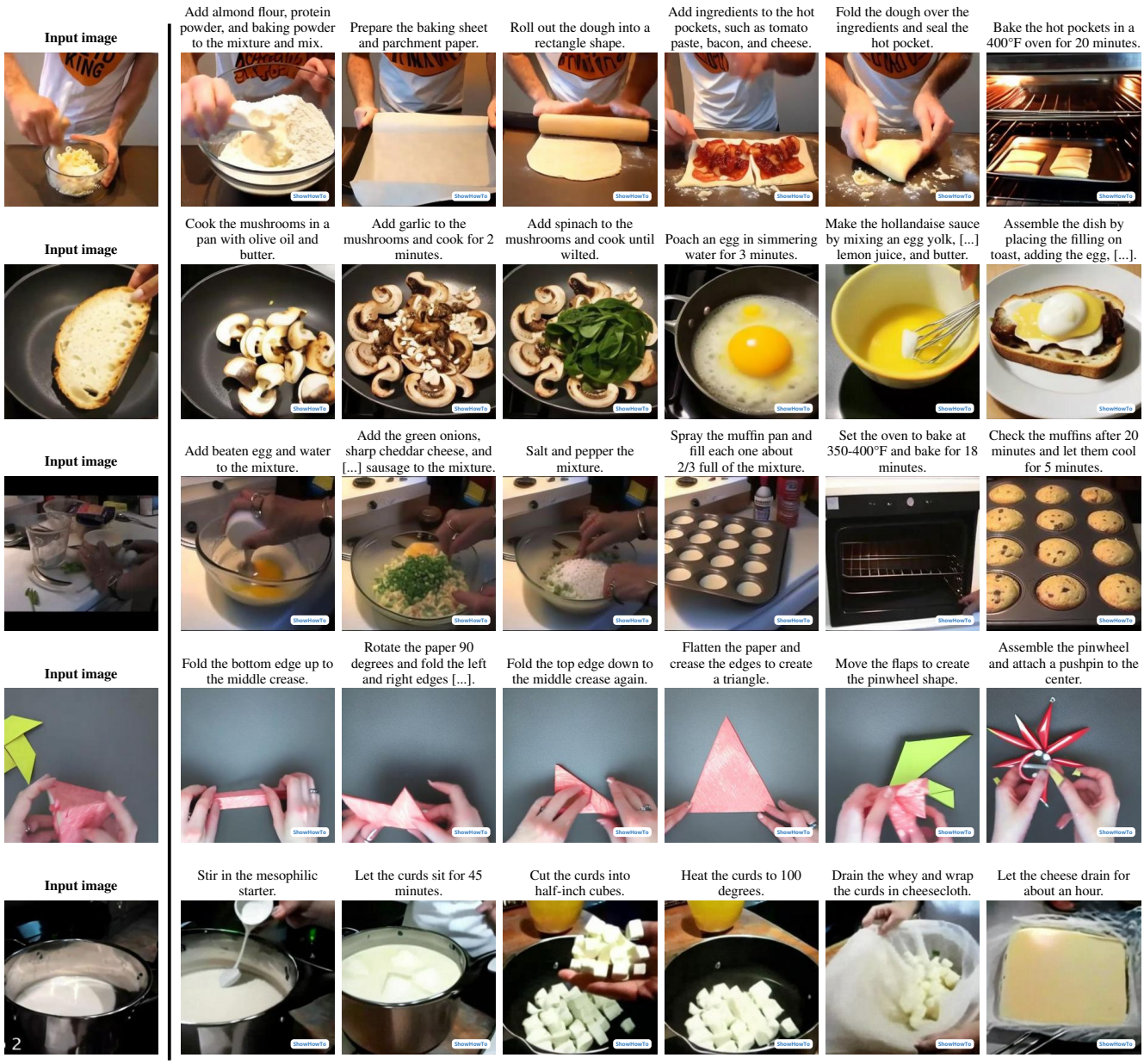

Figure 8. **Additional qualitative results of our method for sequences from the test set.** Given the input image (left) and the textual instructions (top), ShowHowTo generates step-by-step visual instructions while maintaining objects from the input image.

[22] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, 2023. 2

[23] Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, and Josef Sivic. Cross-task weakly supervised learning from instructional videos. In *CVPR*, 2019. 2, 4

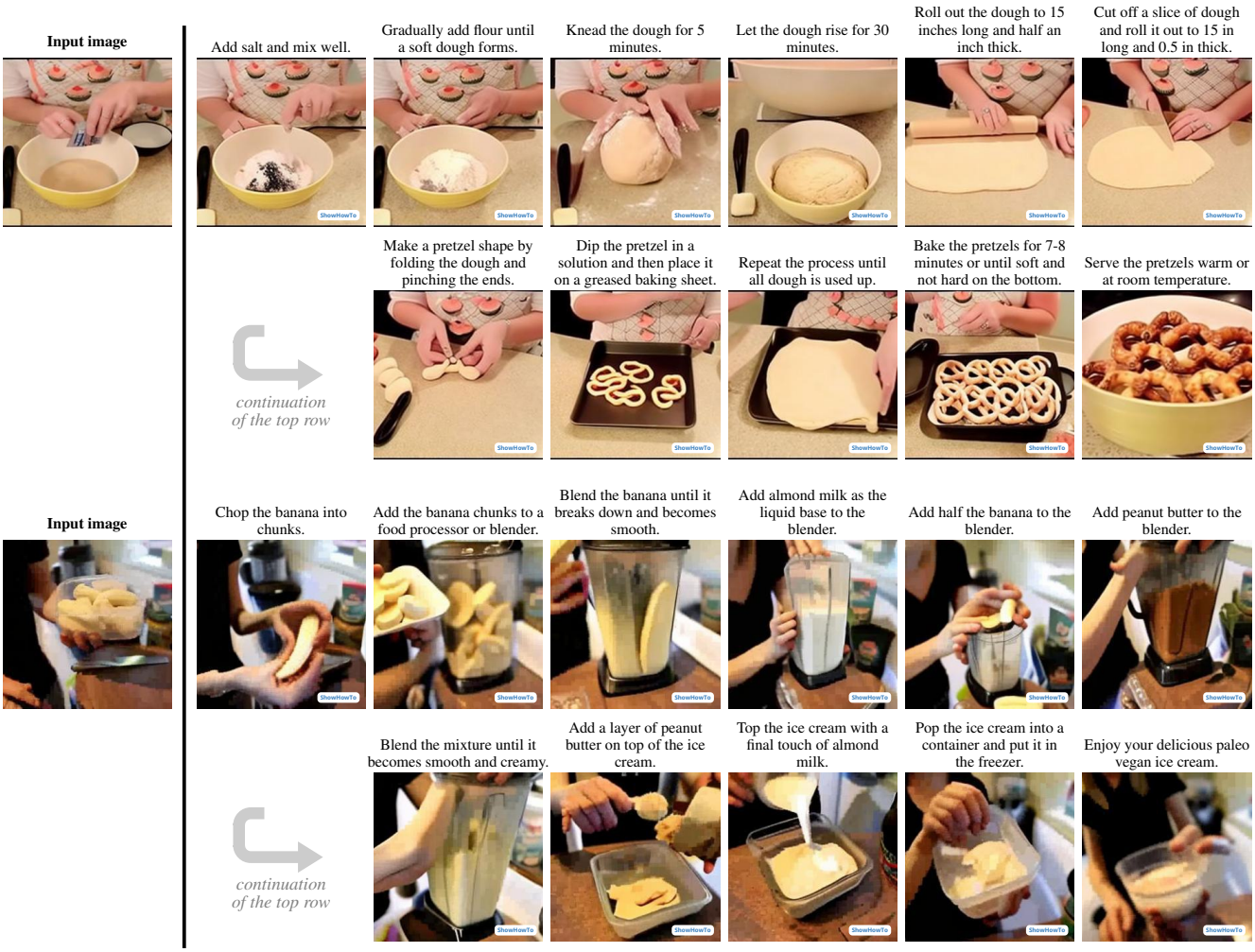

Figure 9. **Qualitative results of our method for sequences from the test set.** Our model can generate long sequences of visual instructions while being consistent with the input image and the text prompts.

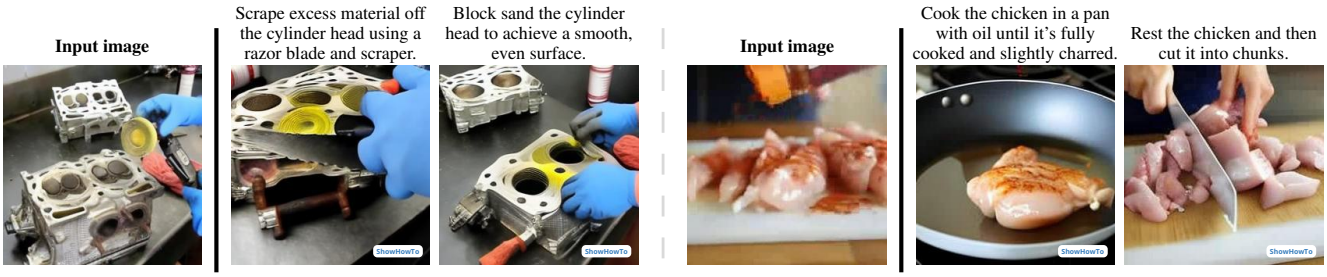

Figure 10. **Failure modes.** The model can struggle with rare objects and tools (left), or it can fail to update object states after state-changing actions (right).

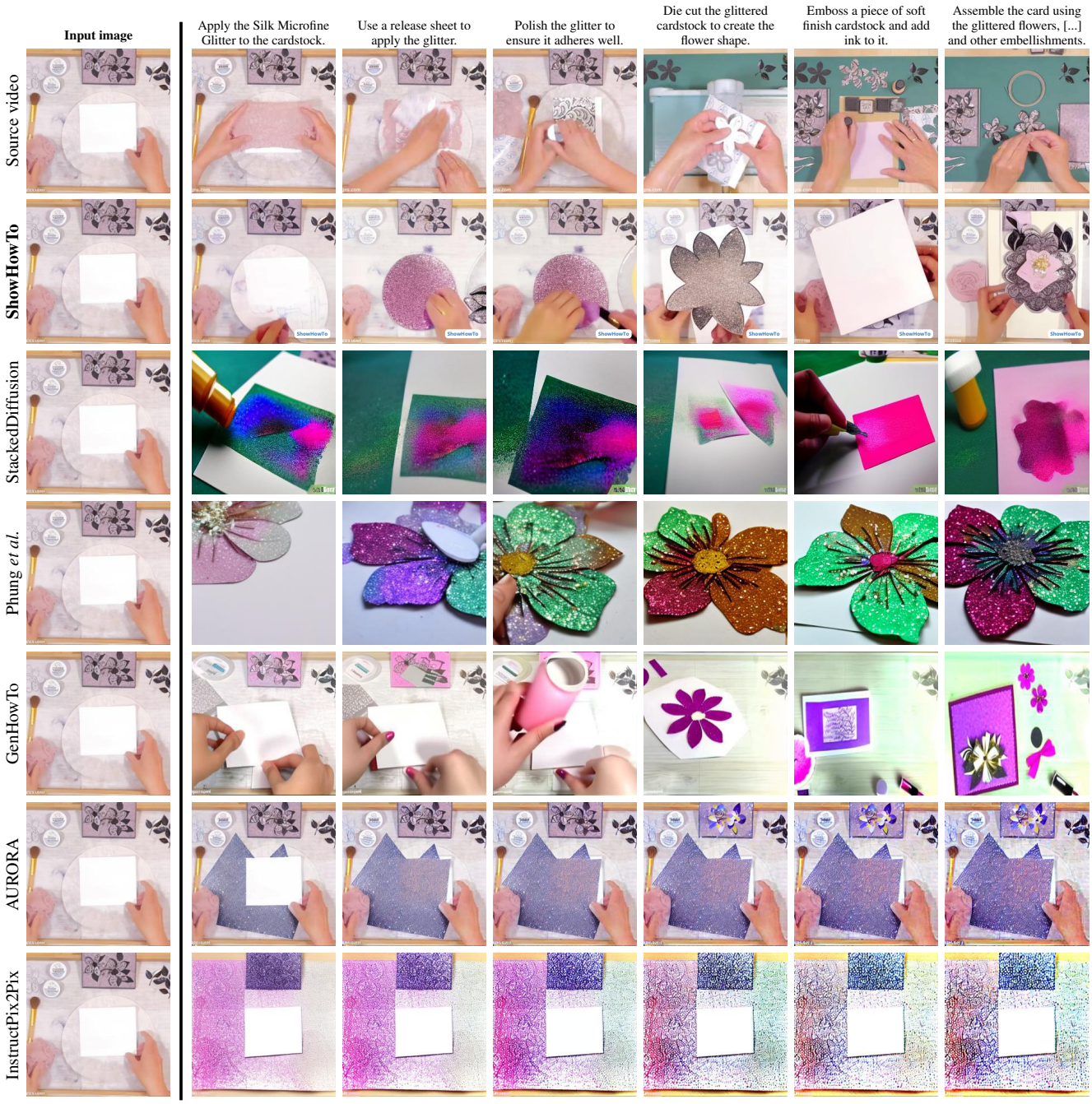

Figure 11. **Additional qualitative comparison** using the input image (left) and the textual instructions (top) for the task of making a cardboard flower with glitter. Only ShowHowTo can produce convincing steps while preserving the input scene.

```
  0.50 -   2.60: Hi, it's Matthew in his pressure cooker again.
  2.60 -   6.09: And today I'm going to make baked potatoes in the
                 pressure cooker.
  6.09 -  11.71: Obviously they're not baked potatoes, just an easy way
                 to make something like a baked potato.
 11.71 -  13.29: It's basically going to be steaming them.
 13.29 -  18.05: So I use my aluminum foil trick.
 18.05 -  19.53: Put some aluminum foil in there.
 19.53 -  27.84: Now a lot of pressure cookers do come with standoffs so
                 that you can put things in where they're sitting above
                 the water.
 28.86 -  29.70: I'm going to make two little rings.
 29.70 -  38.50: I'm going to take my potatoes, put them in so that
                 they're going to be steamed nicely.
 38.50 -  40.42: Because you don't want to actually boil them.
 40.42 -  42.28: You want to steam them.
 42.28 -  44.99: And I've got my cup and a half of water.
 44.99 -  50.23: I'll just dump that in.
 50.23 -  52.27: See what I've got in here?
 52.27 -  53.47: So they're just sitting there.
 53.47 -  54.31: They're above the water.
 54.31 -  57.49: Now, the amount of time varies quite a bit.
 59.85 -  62.05: For these ones I'm going to put them in for about 15
                 minutes.
 62.05 -  63.76: Smaller potatoes maybe a little bit less.
 63.76 -  65.02: Larger potatoes a little bit more.
 65.02 -  68.52: The nice thing about this is it doesn't tie up anything
                 else.
 68.52 -  69.90: It doesn't heat up your kitchen.
 69.90 -  73.88: It's really good for in the summer when you want to have
                 something like a baked potato.
 73.88 -  75.48: It's just going to be steamed.
 75.48 -  76.08: So there it is.
 76.08 -  76.94: 15 minutes.
 76.94 -  80.48: When they're done you just pop them out and eat them
                 like a normal baked potato.
 80.48 -  81.47: I hope you find this useful.
 81.47 -  87.93: If you want to hear more ideas or have any questions
                 leave a comment, send me an email and I'll see what I
                 can do for you.
 89.45 -  92.67: I hope you're enjoying your pressure cooker as much as
                 I'm enjoying mine.
 92.67 -  93.92: Bye.
 93.92 -  96.84: And here we go.
 96.84 - 101.70: I'll pull it out.
101.70 - 104.60: There is the baked potato or steamed potato actually.
104.60 - 107.52: It's nice and it's pretty dry on the outside.
108.55 - 110.37: Just cut it open.
110.37 - 112.43: It's nice and soft.
112.43 - 115.07: Well cooked on the inside.
115.07 - 117.41: Great to cook it up however you're going to make a meal.
117.41 - 119.35: If you're going to eat it like a traditional baked
                 potato.
119.35 - 121.63: If you're going to use it for potato salad or whatever.
121.63 - 123.79: I hope you enjoyed it.
123.79 - 126.75: If you want to see any other ideas, check my channel.
126.75 - 128.35: See what other things I've got posted.
128.35 - 134.16: If you've got ideas that you don't know how to do,
                 send me an email or leave a comment and I'll see what I
                 can do.
134.16 - 134.70: Hope you enjoyed it.
134.70 - 135.04: Bye.
```

```
hi it's math units pressure cooker again
and today I'm going to make baked
potatoes in the pressure cooker
obviously they're not baked potatoes
just an easy way to make something like
a baked potato it's basically going to
be steaming them so I use my aluminum
foil trick yeah it's a moment of oil in
there now a lot of pressure cookers do
come with standoffs but you can cook
eggs in where they're sitting above the
water and make two little rings I'm
going to take my potatoes put them in so
they're going to be steamed nicely
because you don't want to actually boil
them on it you want to see them and go a
cup and half water I'll just dump that

in see see what I've got in here so
they're just sitting there they're above
the water now the amount of time varies
quite a bit for these ones I'm going to
put them in for about 50 minutes smaller

potatoes may be a little bit less larger
potatoes a little bit more a nice thing
about this is it doesn't tie up anything

else doesn't heat up your kitchen it's
really good for in the summer when you
want to have a something like a baked
potato it's just going to be steamed so
there it is 15 minutes when they're done

you just pop them out and eat them like
a normal baked potato I hope you find
this useful if you want to hear more
ideas or have any questions leave a
comment send me an email and I'll I'll
see what I can do here enjoy your I hope
you're enjoying your pressure cooker as
much as I've enjoyed mine bye and here

we go hold it there is Becky or steamed

potato actually placing some pretty dry
on the outside just cut it open it's

nice and nice this off well cooked on
the inside great to cook it up however
you're going to make me know if you need
it like a traditional baked potato we're

going to use it for potato salad or
whatever I hope you enjoyed it if you
want to see any other ideas so, check my
channel see the other things I've got
posted you've got ideas so that you
don't know how to do send me an email or
leave a comment and I'll see what
you do hope you enjoyed it I
```

Figure 12. Comparison between WhisperX [2] speech transcription (left) and YouTube ASR (right) for the same *'How to bake a potato in the pressure cooker'* video. In contrast to YouTube ASR, WhisperX can correctly split the narrations into individual sentences. It also makes significantly fewer errors; for example, it correctly recognizes that the potatoes should be cooked for 15 minutes, not 50 minutes (timestamp 59.85).

Prompt for filtering non-instructional videos

Based on the following video title and partial transcript segment, determine if the video is instructional in nature, where "instructional" means it involves actively demonstrating or teaching how to perform a specific task or activity with physical steps (e.g., cooking a recipe, repairing something, crafting, etc.). Respond with 'Yes' if the video is actively demonstrating or teaching how to perform a specific task, or 'No' if it is not. Then provide a single sentence explanation.

Examples of instructional videos:
• How to Bake a Chocolate Cake
• Repairing a Leaky Faucet
• Learn to Knit a Scarf

Examples of non-instructional videos:
• Discussing Fashion Trends
• Product Reviews and Opinions
• A Vlog of My Daily Life

Example 1:
Video Title: *Red Dead Redemption 2 - Herbert Moon and Strange man Easthereggs In Armadillo [SPOILERS]*
Video Transcript: *"oh you're back I feared the worst it's all here waiting for you who's that I don't know it's just a little portrait somebody gave me once I always quite liked it why no reason just seem familiar anyway this area is closed to the public if you want to shop here you better act right move you long streak of piss who do you think you are for God's sake get out you degenerate you blew it get out of my store if you don't leave there will be problems okay okay stay calm oh you'll (...)"*
Is this video actively demonstrating or teaching how to perform a specific task? No
Explanation: The video is not actively demonstrating or teaching how to perform a specific task; it appears to be showcasing or discussing Easter eggs in the game Red Dead Redemption 2.

Example 2:
Video Title: *Fantastic VEGAN Cupcakes with Raspberry Frosting*
Video Transcript: *"hey there I'm chef Annie and tomorrow is Valentine's Day so we are making some extra special cupcakes for this occasion can you believe that we have not made cupcakes on this channel it's about time so today I'm going to show you how to present these cupcakes so they look impressive and absolutely beautiful so enough copy let's cook it so we're going to start by mixing together our wet ingredients (...)"*
Is this video actively demonstrating or teaching how to perform a specific task? Yes
Explanation: The video actively demonstrates and teaches how to make vegan cupcakes with raspberry frosting, as indicated by the detailed steps and instructions given by the chef.

Example 3:
Video Title: *How To: Piston Ring Install*
Video Transcript: *"hey it's Matt from how to motorcycle repair comm just got done doing a top end on a YZF 250 or yz250 F and I thought I'd do a quick video on how to install a piston ring the easy way now I've done this in the past too but most people will take the ends here and spread it and put it on but you can potentially damage the ring so an easier way to do that is just to take this right here incident in the groove that you need then you bend one up (...)"*
Is this video actively demonstrating or teaching how to perform a specific task? Yes
Explanation: The video is actively demonstrating or teaching how to install a piston ring, which is a specific task.

Example 4:
Video Title: *Best gas weed eater reviews Husqvarna 128DJ with 28cc Cycle Gas Powered String Trimmer*
Video Transcript: *"guys i'm shanley today i'm going to tell you about this straight shaft gas-powered trimmer from husqvarna this trimmer runs on a 28 CC two cycle engine it features 1.1 horsepower and a three-piece crankshaft it also has a smart start system as well as an auto return to stop switch and this trimmer is air purge design for easier starting it has a 17 inch cutting path (...)"*
Is this video actively demonstrating or teaching how to perform a specific task? No
Explanation: This video is reviewing the features of a gas-powered trimmer rather than actively demonstrating or teaching how to use it.

Now, determine if the following video is instructional in nature:
Video Title: {Input video title}
Video Transcript: {Input video transcript}
Is this video actively demonstrating or teaching how to perform a specific task?

Figure 13. Prompt used for filtering non-instructional videos using Llama 3. Transcript excerpts are truncated for clarity, the full prompt will be released with the code.

Prompt for step extraction from an instructional video

Below are transcripts from YouTube instructional videos and their corresponding extracted steps in a clear, third-person, step-by-step format like WikiHow. Each step is concise, actionable, and temporally ordered as they occur in the video. The steps include start and end timestamps indicating when the steps are carried out in the video. Follow this format to extract and summarize the key steps from the provided transcript.

Example 1:

YouTube Video Title: "*BÁNH TÁO MINI - How To Make Apple Turnovers — Episode 11 — Taste From Home*"

YouTube Video Transcript:
00.87 - 07.79: "*Hey little muffins, today we will make together a super easy, quick and delicious apple turnovers.*"
07.79 - 09.35: "*40 minutes for all the process.*"
09.35 - 11.95: "*Seriously, can someone deny them?*"
11.95 - 13.63: "*Ok, let's begin.*"
13.63 - 18.82: "*First of all, combine the apple cubes, lemon juice, cinnamon and sugar in a bowl.*"
26.69 - 29.59: "*Mixing, mixing, mixing.*"
29.59 - 32.62: "*Apple and cinnamon always go perfectly together.*"
32.62 - 43.52: "*Now using a round cutter or glass like me, cut 15 rounds from the pastry sheet.*"
57.86 - 64.99: "*Here comes the fun part.*"
64.99 - 69.97: "*Spoon about 2 teaspoons apple mixture in the center of one round.*"
69.97 - 74.41: "*Using your fingers, gently fold the pastry over to enclose filling.*"
88.47 - 104.48: "*After that, use a fork and press around the edges to seal and make your apple turnovers look more beautiful.*"
104.48 - 105.84: "*This is how it looks like.*"
109.99 - 113.53: "*I will show you one more time to make sure that you understand the technique.*"
113.53 - 117.20: "*And if you still find my apple turnovers too ugly, I'm really sorry.*"
(...)

Extracted Steps:

```
[{ "WikiHow Title": "How to Make Apple Turnovers" },
  { "steps": [
    { "step": 1, "instruction": "Combine apple cubes, lemon juice, cinnamon, and sugar in a bowl
        .", "start_timestamp": 13.63, "end_timestamp": 18.82 },
    { "step": 2, "instruction": "Mix the ingredients thoroughly.", "start_timestamp": 26.69, "
        end_timestamp": 29.59 },
    { "step": 3, "instruction": "Cut 15 rounds from the pastry sheet using a round cutter or a
        glass.", "start_timestamp": 32.62, "end_timestamp": 43.52 },
    { "step": 4, "instruction": "Spoon about 2 teaspoons of the apple mixture into the center of
        one round.", "start_timestamp": 64.99, "end_timestamp": 69.97 },
    { "step": 5, "instruction": "Gently fold the pastry over to enclose the filling using your
        fingers.", "start_timestamp": 69.97, "end_timestamp": 74.41 },
    { "step": 6, "instruction": "Press around the edges with a fork to seal and beautify the
        turnovers.", "start_timestamp": 88.47, "end_timestamp": 104.48 },
    { "step": 7, "instruction": "Repeat the technique until all turnovers are formed.", "
        start_timestamp": 109.99, "end_timestamp": 113.53 },
    { "step": 8, "instruction": "Lightly beat one egg in a small bowl.", "start_timestamp":
        151.62, "end_timestamp": 157.46 },
    { "step": 9, "instruction": "Egg wash the apple turnovers to give them a gorgeous light brown
        color after baking.", "start_timestamp": 157.46, "end_timestamp": 164.10 },
    { "step": 10, "instruction": "Bake the apple turnovers at 180°C for 18-20 minutes until golden
        .", "start_timestamp": 164.10, "end_timestamp": 174.87 },
    { "step": 11, "instruction": "Enjoy the freshly baked apple turnovers.", "start_timestamp":
        178.17, "end_timestamp": 185.55 }]}]
```

Example 2: (...)

Now, extract the steps from the following transcript:

YouTube Video Title: {Input video title}

YouTube Video Transcript: {Input video transcript}

Extracted Steps:

Figure 14. Prompt used for generating instructional steps with start-end timestamps using Llama 3. The prompt is truncated for clarity, the full prompt will be released with the code.