Stitch-a-Demo: Creating Video Demonstrations from Multistep Descriptions

Supplementary Material

A. Supplementary video

We provide a supplementary video that shows an overview of the paper. It also shows qualitative video examples that best illustrates the visual demonstration we obtain from multistep descriptions.

B. Stitch-a-Demo dataset

LLM prompts. As discussed in Sec. 3.3, we create weakly supervised training data $\mathcal{D}_w$ for our task. We use a recent language model, Llama 3.1 70B [14], to extract step descriptions from the narrations, and novel procedure combinations. We show the LLM prompt for each of the two tasks below:

Firstly, we provide narrations and ask the language model to provide the step descriptions. Here are the prompts for cooking, woodworking, and gardening:

> **System:** Help summarize the steps of this recipe whose narrations with timestamps are given. Timestamp is given in seconds.
> **User:** Given the narrations and the timestamp of a video in the format '[start_time-end_time] narration text', tell the recipe being made in this video and list down the steps required to complete this recipe. For each step, list down the timestamp of the corresponding narrations that best describe the step. Do not list the introduction, explanation, or comment as a step. Answer in this format: 'Recipe: Name of the recipe and brief detail.
> Step 1: [start_time-end_time] description of the step
> Step 2: [start_time-end_time] description of the step and so on.'. Here are narrations:
> NARRATION COMES HERE
> **Assistant:**

> **System:** Help summarize the steps of this woodworking project whose narrations with timestamps are given. Timestamp is given in seconds.
> **User:** Given the narrations and the timestamp of a video in the format '[start_time-end_time] narration text', tell the woodworking project being made in this video and list down the steps required to complete this project. For each step, list down the timestamp of the corresponding narrations that best describe the step. Do not list the introduction, explanation, or comment as a step. Answer in this format: 'Project: Name of the project and brief detail.
> Step 1: [start_time-end_time] description of the step
> Step 2: [start_time-end_time] description of the step and so on.'. Here are narrations:
> NARRATION COMES HERE
> **Assistant:**

> **System:** Help summarize the steps of this gardening project whose narrations with timestamps are given. Timestamp is given in seconds.
> **User:** Given the narrations and the timestamp of a video in the format '[start_time-end_time] narration text', tell the gardening project being made in this video and list down the steps required to complete this project. For each step, list down the timestamp of the corresponding narrations that best describe the step. Do not list the introduction, explanation, or comment as a step. Answer in this format: 'Project: Name of the project and brief detail.
> Step 1: [start_time-end_time] description of the step
> Step 2: [start_time-end_time] description of the step and so on.'. Here are narrations:
> NARRATION COMES HERE
> **Assistant:**

After obtaining the summaries, we feed $N = 2/3/4$ procedures from similar tasks to the language model and ask it to create a novel procedure, following all the desired constraints of realism and correctness. We show the prompt for $N = 3$ and for three different domains respectively here:

> **System:** Generate a new recipe by combining steps from different recipes.
> **User:** You are tasked with creating a new recipe by combining steps from 3 provided recipe summaries. Your goal is to seamlessly integrate steps from each recipe, switching between them only when necessary due to differences in ingredients, techniques, or tools. Ensure that you do not introduce any new ingredients or steps beyond what is outlined in the summaries. Also, make sure to use at least one step from all recipes. Answer 'Not Possible' if this is not possible. Format your response as follows:
> Step 1 (Step _ in Recipe _): [Description of the step]
> Step 2 (Step _ in Recipe _): [Description of the step]
> ...
> Step t (Step _ in Recipe _): [Description of the step]

Explanation: [Any explanation that you want to provide]
Here are the procedures:
RECIPE 1 COMES HERE
RECIPE 2 COMES HERE
RECIPE 3 COMES HERE
Assistant:

System: Generate a new woodworking plan by selecting steps from different project instructions.
User: You are tasked with creating a new woodworking plan by selecting steps from 3 provided project summaries. Your goal is to seamlessly integrate steps from each plan, switching between them only when necessary due to differences in materials, joinery methods, or tools. Ensure that you do not introduce any new materials, techniques, or steps beyond what is outlined in the summaries. Also, make sure to use at least one step from all project summaries. Answer 'Not Possible' if this cannot be done. Format your response as follows:
Step 1 (Step _ in Project _): [Description of the step]
Step 2 (Step _ in Project _): [Description of the step]
...
Step t (Step t in Project _): [Description of the step]
Explanation: [Any explanation that you want to provide]
Here are the projects:
PROJECT 1 COMES HERE
PROJECT 2 COMES HERE
PROJECT 3 COMES HERE
Assistant:

System: Generate a new gardening plan by selecting steps from different project instructions.
User: You are tasked with creating a new gardening plan by selecting steps from 3 provided project summaries. Your goal is to seamlessly integrate steps from each plan, switching between them only when necessary due to differences in plants, soil preparation methods, or tools. Ensure that you do not introduce any new plants, techniques, or steps beyond what is outlined in the summaries. Also, make sure to use at least one step from all project summaries. Answer 'Not Possible' if this cannot be done. Format your response as follows:
Step 1 (Step _ in Project _): [Description of the step]
Step 2 (Step _ in Project _): [Description of the step]
...
Step t (Step t in Project _): [Description of the step]
Explanation: [Any explanation that you want to provide]
Here are the projects:
PROJECT 1 COMES HERE
PROJECT 2 COMES HERE
PROJECT 3 COMES HERE
Assistant:

Overall, our language model prompts ensure good-quality weakly supervised data; the effectiveness is also shown in Sec. 4.

Stitch-a-Demo-VD dataset curation. In VidDetours's [6] manual annotation, for two videos $V_1, V_2 \in \mathcal{C}$, the annotation specifies a time t in V_1 where a user asks Q to take a detour to time window $[t_2, t_3]$ in V_2. Q is a natural language question like *"Can we substitute garlic with shallots?"*. We use Q, t, and $[t_2, t_3]$ to choose procedure steps from V_1 and V_2. Since there exists a procedure step r' after time t_2 in V_2 that answers Q, ensuring $r' \in R$ implies $v_i \in V_1$ and $v_{i+1} \in V_2$ for some i. This process can be extended to create a procedure with ground truth containing more videos. Note that the dataset has many distinct annotations with both $V_1 \to V_2$ and $V_2 \to V_1$ detours, thus reducing the number of distinct videos in a sample in $\mathcal{D}_d$, see Sec. 3.4 for statistics. To ensure the correctness of the procedures in $\mathcal{D}_d$, we manually verified text-visual step alignment, object-state consistency, and that the video shows the target procedure. This process leverages the annotation effort in [6] to create a clean testing set, with minimal manual verification efforts.

Note that we cannot use the evaluation dataset in Recipe2Video [61] (RecipeQA [74] and Tasty videos [53]) since they do not have ground truth annotations; they compare heuristics like abrupt info gain and visual relevance.

Dataset statistics. We provide more detailed statistics for our datasets across three domains. Tab. 3 shows the number of videos in each dataset. Using these videos, Stitch-a-Demo-MC contains 105542, 100, and 100 samples for cooking, woodworking, and gardening. For $\mathcal{D}_w$, it consists of 444823, 900, and 900 samples for the three domains, respectively. Across different test sets, an average of 53.1 videos with similar content are considered for each sequence, making the test challenging.

We also compare Stitch-a-Demo-MC and Stitch-a-Demo-VD to standard video for object-state consistency using VBench [28] subject consistency at keystep transitions; the curated test sets score only 1.5% and 1% lower than standard video, respectively, showing comparable quality.

C. Results and additional ablations

Baseline implementation details. We use the implementation provided in [25] to extract InternVideo [66] visual

Dataset	Cooking	Woodworking	Gardening
$\mathcal{D}_w$	321139	1012	1026
SaD-MC	2657	100	100
SaD-VD	235	0	0
HT-Step [2]	1087	0	0
COIN [58]	337	50	70
CrossTask [84]	812	130	0

Table 3. Number of videos per dataset for cooking, gardening, and woodworking.

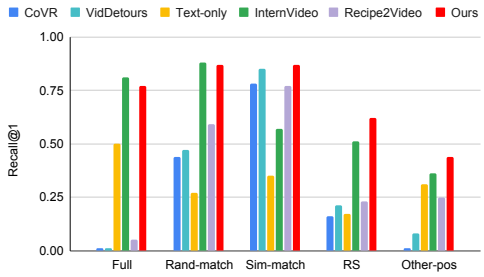

Figure 6. **Result on distractor set splits**. Our model performs competitively on all splits—particularly the more challening RS, Other-pos, and Sim-match.

features. We use the author's provided code for VidDetours [6] and CoVR [62], and retrain on our dataset for a fair comparison. We re-implement Recipe2Video [61] since there is no publicly available codebase. We matched our re-implementation on RecipeQA [74] and obtain a visual relevance of 0.82, better than the reported 0.8, establishing the correctness of our usage.

Creating the distractor set for evaluation. In Sec. 4, we state that we create a *distractor set* containing 499 negative samples for retrieval performance evaluation. We describe the composition of this challenging negative set, that assesses various aspects of the visual demonstration creation. The dataset consists of equal samples from the following strategies:

- Reduced search space top-K samples (RS): We use the set cover algorithm, as introduced in Sec. 3.2, to create a set of challenging options. These options contain visual demonstrations that combine multiple source videos.
- Full videos (Full): We sample videos from the same task that serves as an unmixed distractor in the candidate set.
- Other positives (Other-pos): These are ground truth visual demonstrations for other procedures in the test set.
- Random mix-n-match (Rand-match): These distractors are a random combination of video clips from the same task.
- Mix-n-match w/ similarity (Sim-match): For every step description, we choose the highest similarity visual demonstrations and create a sequence from top retrievals, based on the similarity scores. Note that these options do not consider the visual continuity, unlike our training positives.

Overall, we carefully design negatives in the distractor set that serve as a competitive benchmark for this task.

Additional result splits. We provide performance splits for results presented in Sec. 4. We show results for the following splits:

- Performance split between COIN and CrossTask: Tab. 4 shows performance split between COIN [58] and CrossTask [84]. We show the combined performance in Sec. 4 since both the datasets evaluate the same aspects of the model. We outperform all the baselines on all for metrics for both the datasets. Note that CrossTask [84] does not have gardening videos.
- Performance across various distractor set components: In Sec. C, we introduce the distractor set components. We evaluate the performance with each component of the distractor set. For example, we evaluate the retrieval performance with 99 negative samples from 'Random mix-n-match' and one ground truth. Fig. 6 shows the results. Our method consistently performs good on all distractor sets. In particular, our method outperforms all the competing methods on more difficult distractor sets, *i.e.*, RS, Sim-match, and Other-pos.

Ablations. We present two key ablations:

Weakly-supervised dataset creation: In Sec. 3.3, we introduce a pipeline to create weakly-supervised training data $\mathcal{D}_w$. This method used an LLM to create novel procedures, and the corresponding video demonstrations. We compare the effectiveness of this data augmentation, compared to (a) w/o augmentation, where we train only with videos in $\mathcal{C}$ (*i.e.* HowTo100M [44]) directly, and (b) temporally-sampled procedure combination, where we use videos from the same task and create new procedures by combining steps happening at the same time. The second method also creates novel procedures, however, unlike our approach, it does not leverage an LLM to create *plausible* transitions. Tab. 5 shows the results on Stitch-a-Demo-VD test set. Our method clearly outperforms both the data creation alternatives, showing the effectiveness of our LLM-aided weakly supervised set $\mathcal{D}_w$. We see the same trend in all the test datasets.

Negative sampling strategy: We introduce negative sampling strategies in Sec. 3.2. We compare the effect of individual negative samples in Tab. 5 (bottom). We observe that all three of *step correctness*, *visual continuity*, and *object state continuity* are crucial for our method's performance.

More fine-grained analysis. As in B, we leverage VBench [28] subject consistency at keypoint transitions as a fine-grained measure of procedure retrieval quality. Our method scores 0.67, outperforming the strongest baseline, InternVideo [67], which scores 0.62, demonstrating better object-state consistency in the retrieved clips.

Method	Cooking				Woodworking				Gardening	
	COIN		Crosstask		COIN		Crosstask		COIN	
	MR↓	R@50↑	MR↓	R@50↑	MR↓	R@50↑	MR↓	R@50↑	MR↓	R@50↑
CoVR [62]	95	0.25	97	0.25	33	0.82	40	0.71	26	0.30
VidDetours [6]	35	0.62	38	0.60	46	0.54	44	0.61	40	0.24
Text only	80	0.36	78	0.35	31	0.76	37	0.87	58	0.32
InternVideo [66]	14	0.68	23	0.66	46	0.68	40	0.76	45	0.28
Recipe2Video [61]	25	0.70	27	0.67	75	0.06	76	0.07	76	0.12
Ours	**4**	**0.86**	**4**	**0.91**	**28**	**0.92**	**35**	**0.87**	**25**	**0.36**

Table 4. **Additional results on visual demonstration retrieval.** Comparison of the performance of our method against baselines and prior work for COIN [58] and CrossTask [84]. This table shows per dataset performance for the consolidated results in Tab. 1.

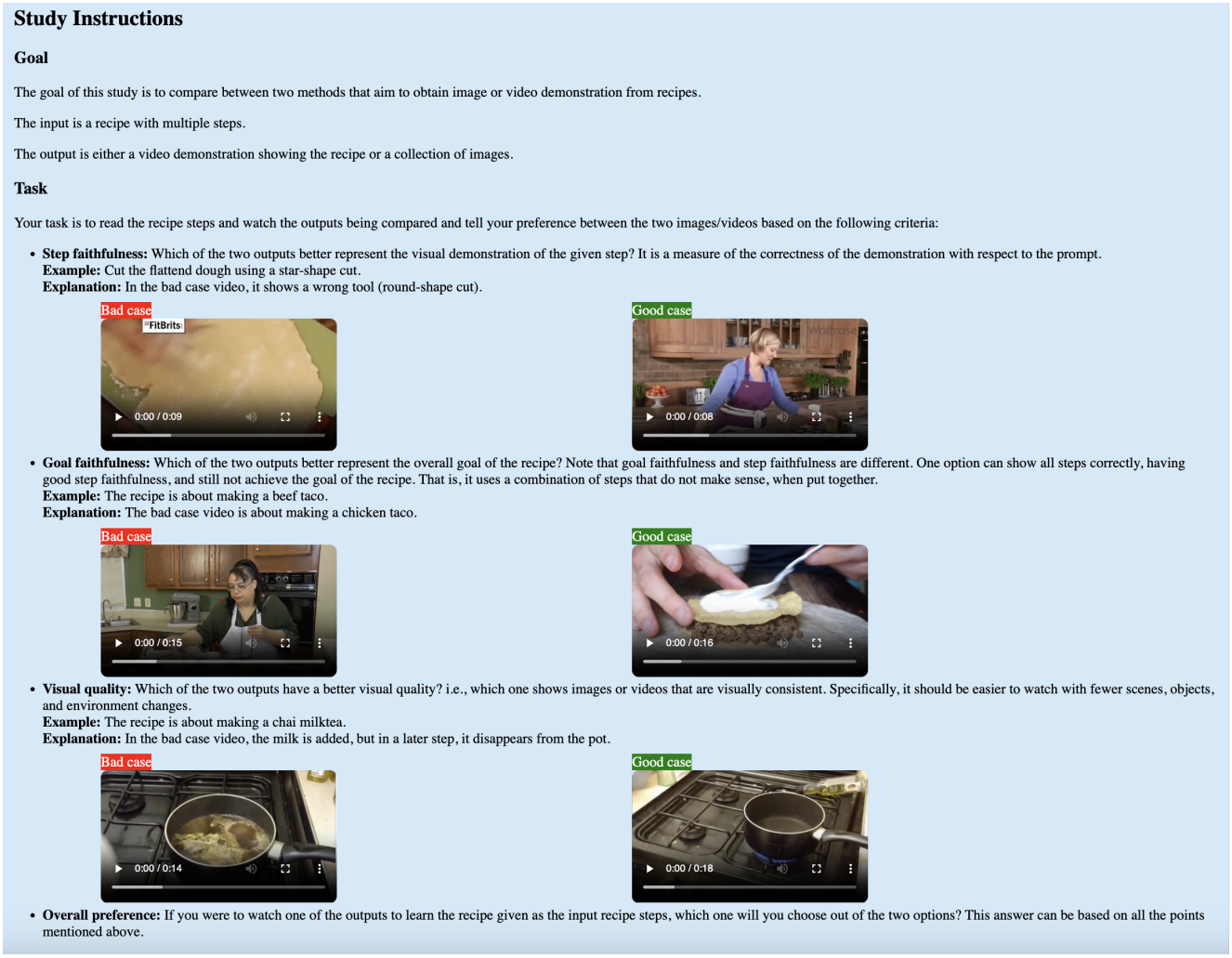

Figure 7. **Human preference study interface instructions.** We provide examples of all axes for human preference study—step faithfulness, goal faithfulness, visual quality and the overall preference.

D. Adaptive search space reduction

In Sec. 3.2 we introduce our set cover algorithm that reduces the search space for practical deployment. For computational complexity, given a procedure query with M steps and a pool of N videos with K clips each, our method selects the top S clips per step in $O(NK \log S)$ time and explores coherent combinations in $O(SM^2)$, finishing at combinations that minimize switches between sources.

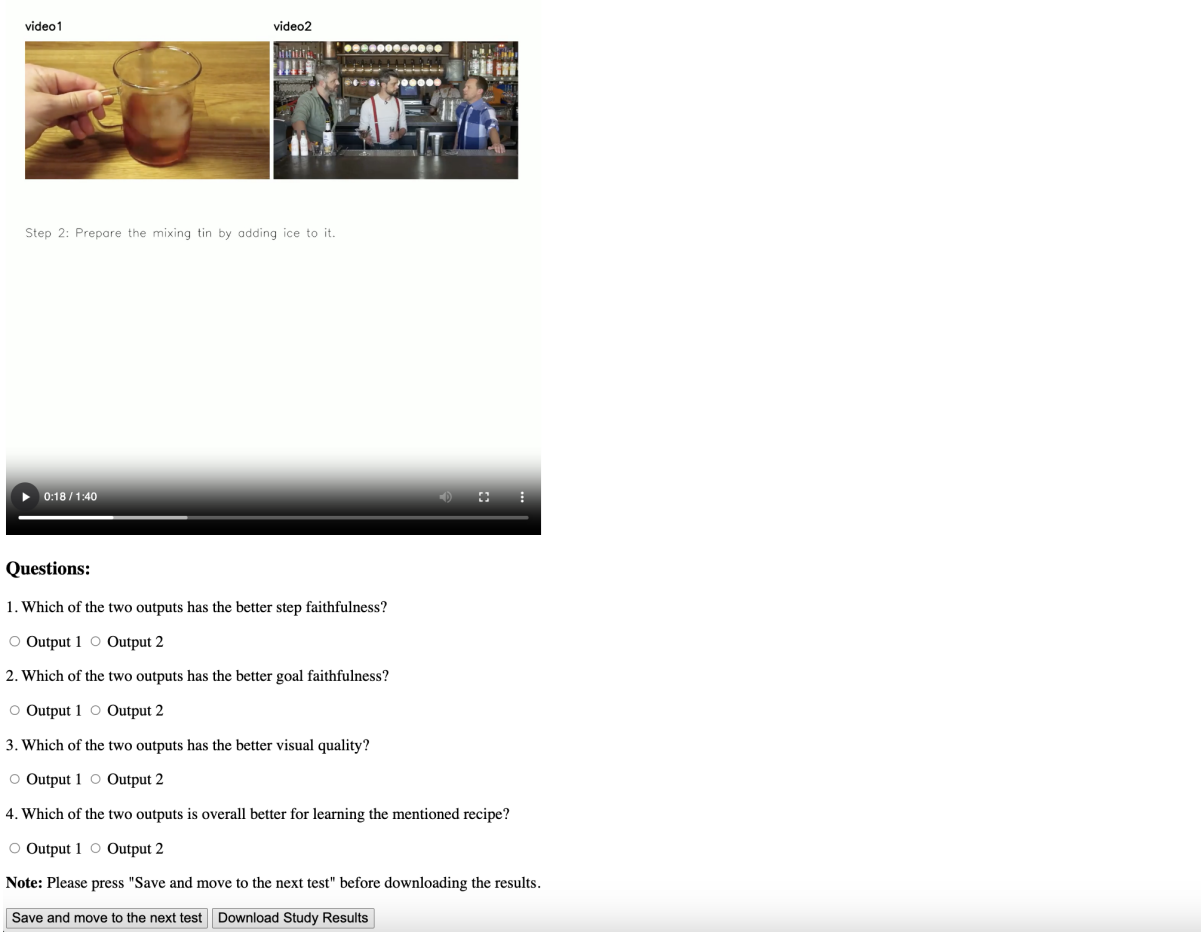

Figure 8. **Human preference study submission form.** The video in the interface shows both the candidate procedures side by side, and the step description is shown below. The video is followed by four questions, asking about each axis, and the result is saved as a CSV file.

Method	MedR↓	R@1↑	R@5↑
w/o augmentation	70	0.03	0.14
Temporally-sampled procedures	10	0.18	0.42
Weakly supervised $\mathcal{D}_w$ (ours)	**3.5**	**0.23**	**0.56**

Cor	Con	OSC	MedR↓	R@1↑	R@5↑
✓			9	0.17	0.41
	✓		171	0	0.03
		✓	11	0.11	0.39
✓	✓		9	0.18	0.45
	✓	✓	15	0.07	0.21
✓		✓	**5**	0.15	**0.54**
✓	✓	✓	**5**	**0.22**	0.52

Table 5. **Ablations.** Our weakly-supervised data is effective for training, compared to alternatives (top). All the constraints for hard negatives are useful (bottom). (Cor: step/goal correctness, Con: visual continuity, OSC: object state consistency.)

E. Human preference study interface

We invited 20 annotators for evaluation, with each comparison annotated by three people. The annotators show high consistency (85.2% pairwise agreement). Fig. 7 and Fig. 8 show our designed human preference study interface instructions and questions, respectively. We first provide instructions, followed by examples of all aspects we evaluate—step faithfulness, goal faithfulness, visual quality and overall preference (Fig. 7). Step faithfulness evaluates which of the two methods shows correct video clips for a given step. Goal faithfulness takes it a step further by assessing if the overall goal of the procedure R is satisfied or not. Visual quality checks for which of the options is easier to watch (with fewer context jumps) and finally, overall preference captures which output the user would prefer. The two options are randomly shuffled to avoid any bias in the human subjects. All the subjects are unrelated to this project and briefed about the task before the study.