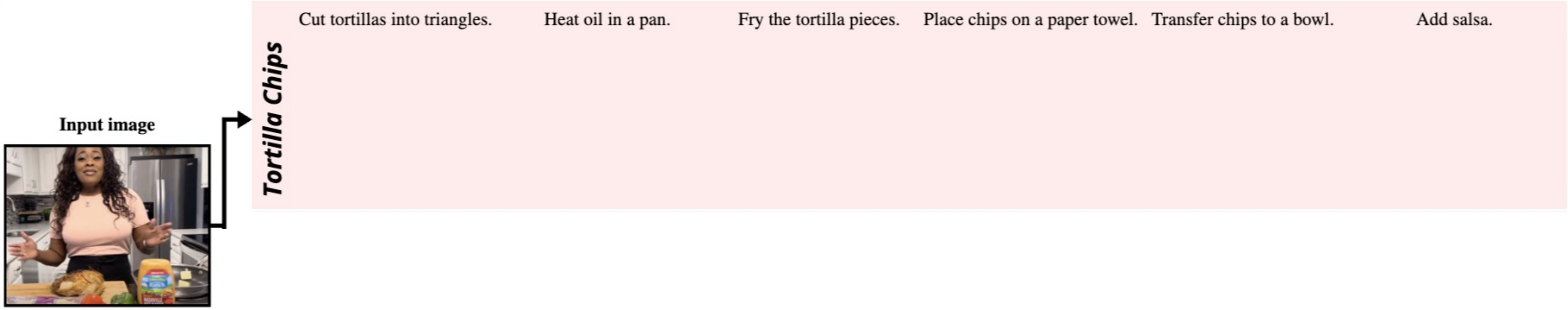
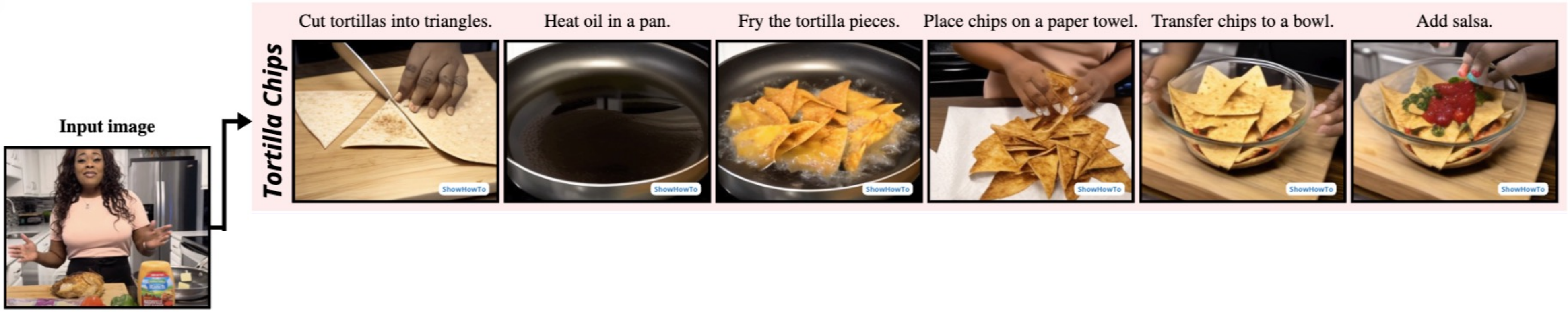
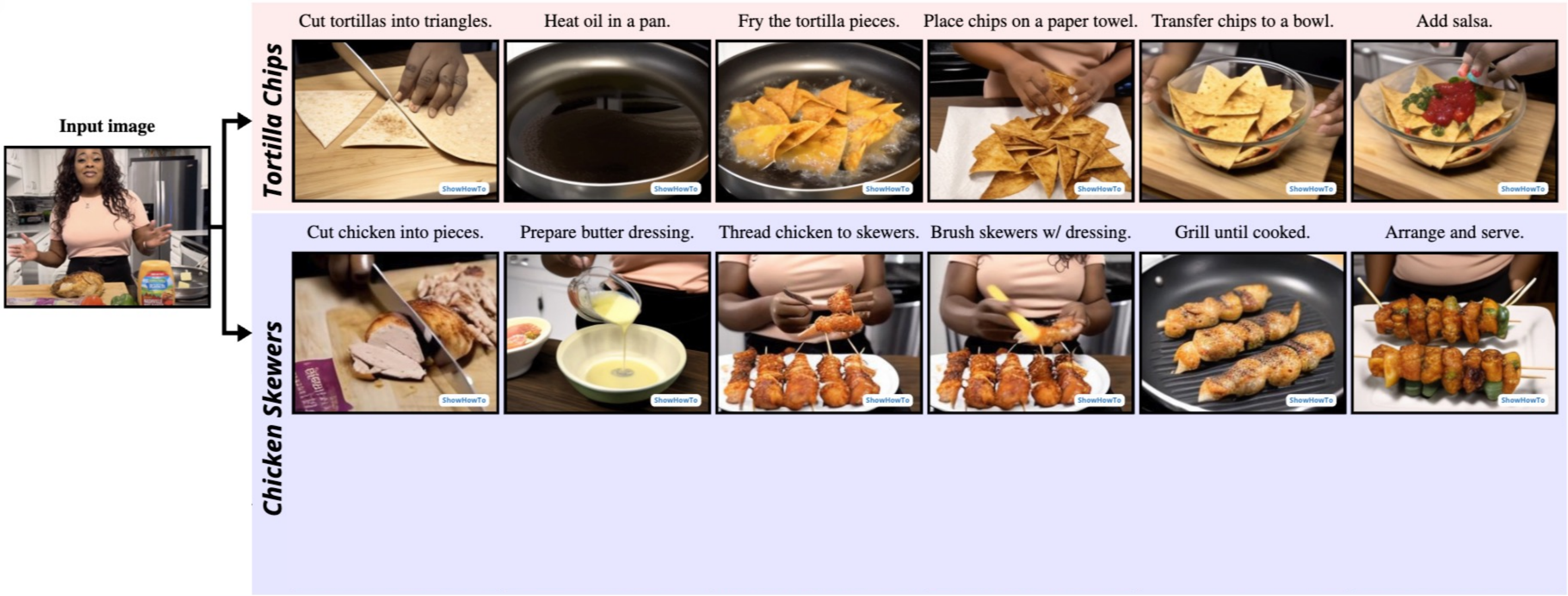
ShowHowTo: Generating Scene-Conditioned Step-by-Step Visual Instructions

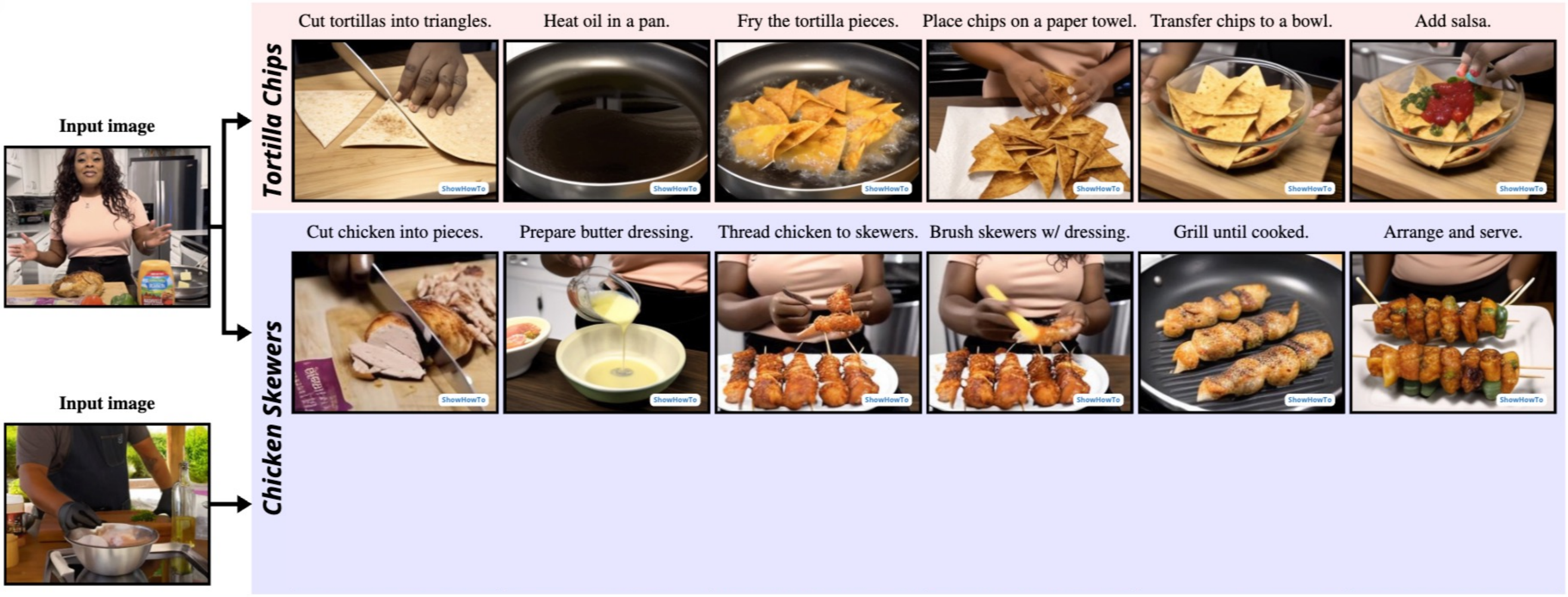
Tomáš Souček[1] Prajwal Gatti[2] Michael Wray[2] Ivan Laptev[3] Dima Damen[2] Josef Sivic[1]

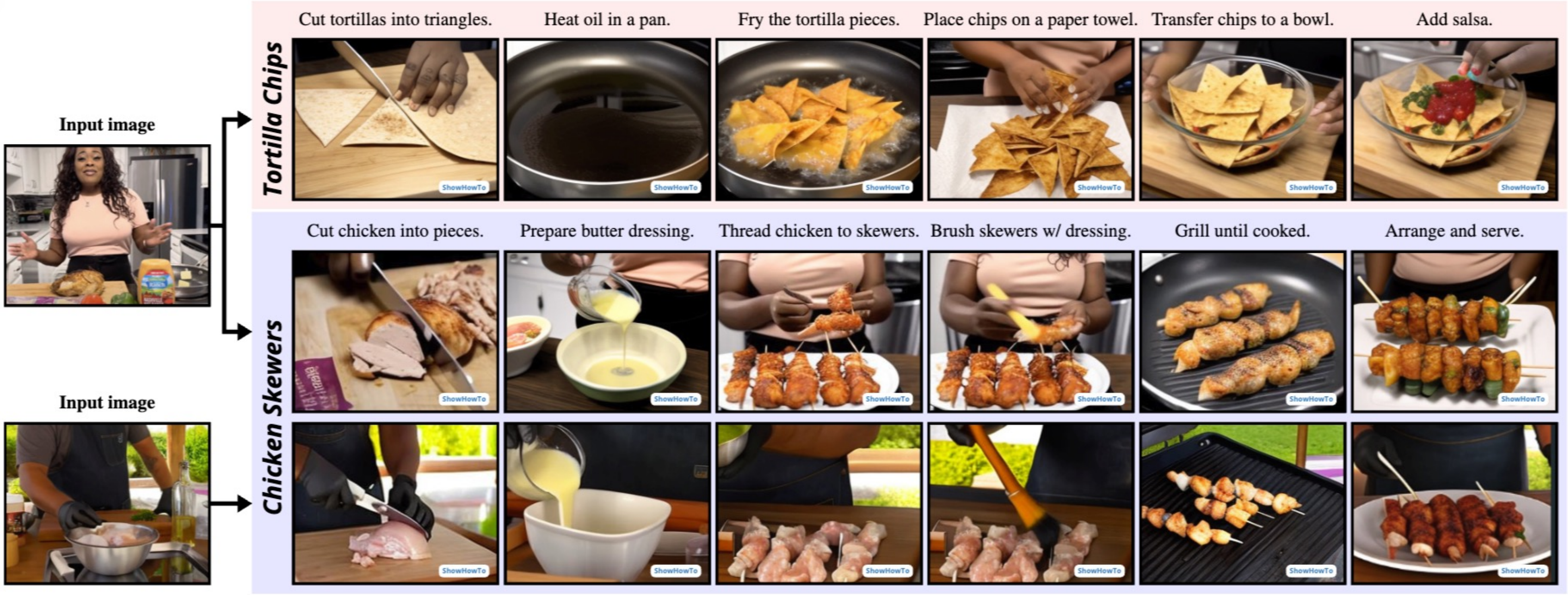
[1]CIIRC CTU [2]University of Bristol [3]MBZUAI


Input image
Tortilla Chips
Cut tortillas into triangles.
Heat oil in a pan.
Fry the tortilla pieces.
Place chips on a paper towel.
Transfer chips to a bowl.
Add salsa.


Input image
Tortilla Chips
Cut tortillas into triangles.
Heat oil in a pan.
Fry the tortilla pieces.
Place chips on a paper towel.
Transfer chips to a bowl.
Add salsa.
ShowHowTo


Input image
Tortilla Chips
Cut tortillas into triangles.
Heat oil in a pan.
Fry the tortilla pieces.
Place chips on a paper towel.
Transfer chips to a bowl.
Add salsa.
Chicken Skewers
Cut chicken into pieces.
Prepare butter dressing.
Thread chicken to skewers.
Brush skewers w/ dressing.
Grill until cooked.
Arrange and serve.


Input image
Input image
Tortilla Chips
Cut tortillas into triangles.
Heat oil in a pan.
Fry the tortilla pieces.
Place chips on a paper towel.
Transfer chips to a bowl.
Add salsa.
Chicken Skewers
Cut chicken into pieces.
Prepare butter dressing.
Thread chicken to skewers.
Brush skewers w/ dressing.
Grill until cooked.
Arrange and serve.
ShowHowTo


Input image
Input image
Tortilla Chips
Cut tortillas into triangles.
Heat oil in a pan.
Fry the tortilla pieces.
Place chips on a paper towel.
Transfer chips to a bowl.
Add salsa.
Chicken Skewers
Cut chicken into pieces.
Prepare butter dressing.
Thread chicken to skewers.
Brush skewers w/ dressing.
Grill until cooked.
Arrange and serve.
ShowHowTo

Related Work

“A man hiking in the mountains with a backpack"

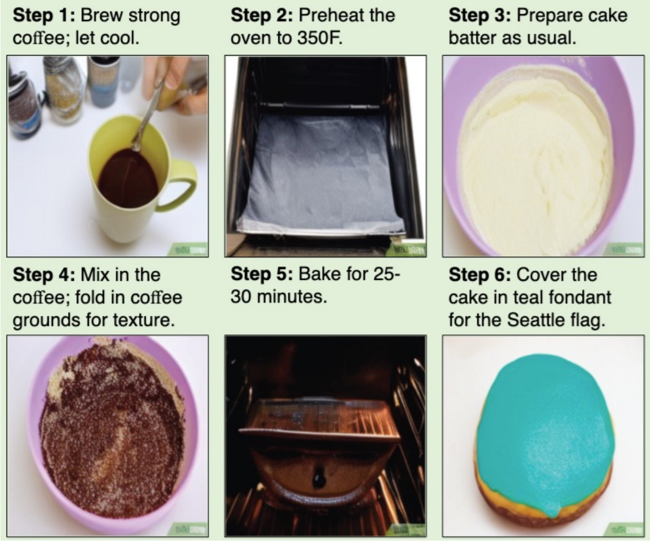
Image to Video Generation

DynamicCrafter (Xing et al ECCV’24),
Stable Video Diffusion (Blattmann et al arXiv’23),
CogVideo (Hong et al ICLR‘23)

Editing Object States or Actions

GenHowTo (Soucek et al CVPR’24),
AURORA (Krojer et al NeurIPS’24)

Baking a cake



Step 1: Brew strong coffee; let cool.

Step 2: Preheat the oven to 350F.

Step 3: Prepare cake batter as usual.

Step 4: Mix in the coffee; fold in coffee grounds for texture.

Step 5: Bake for 25-30 minutes.

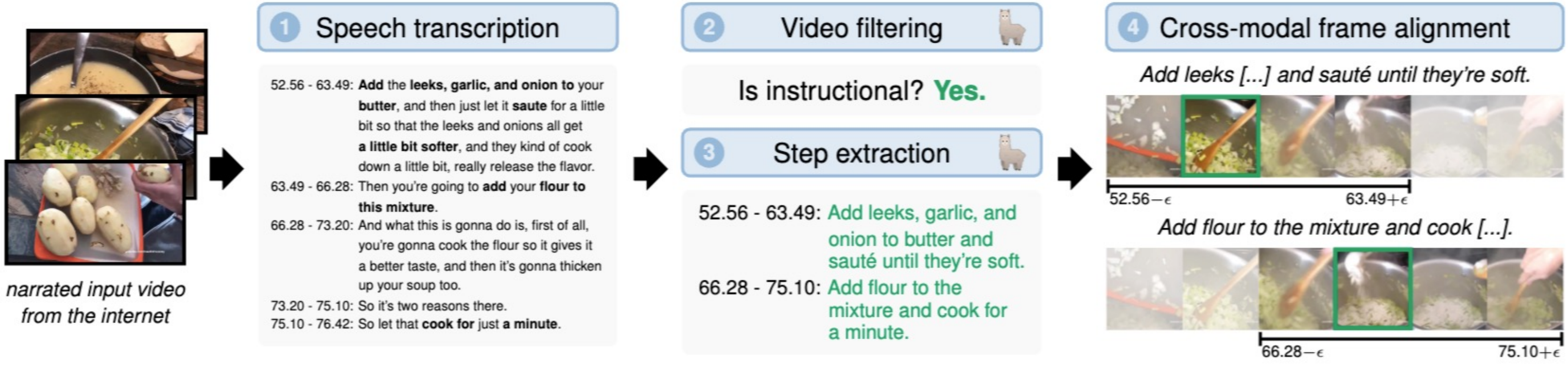
Step 6: Cover the cake in teal fondant for the Seattle flag.

Instruction Generation

StackedDiffusion (Menon et al CVPR’24)

Building a Large-Scale ShowHowTo Dataset

Building a Large-Scale ShowHowTo Dataset



narrated input video from the internet

1 Speech transcription

52.56 - 63.49: **Add** the **leeks, garlic, and onion to** your **butter**, and then just let it **saute** for a little bit so that the leeks and onions all get **a little bit softer**, and they kind of cook down a little bit, really release the flavor.
63.49 - 66.28: Then you're going to **add** your **flour to this mixture**.
66.28 - 73.20: And what this is gonna do is, first of all, you're gonna cook the flour so it gives it a better taste, and then it's gonna thicken up your soup too.
73.20 - 75.10: So it's two reasons there.
75.10 - 76.42: So let that **cook for** just **a minute**.

2 Video filtering

Is instructional? **Yes.**

3 Step extraction

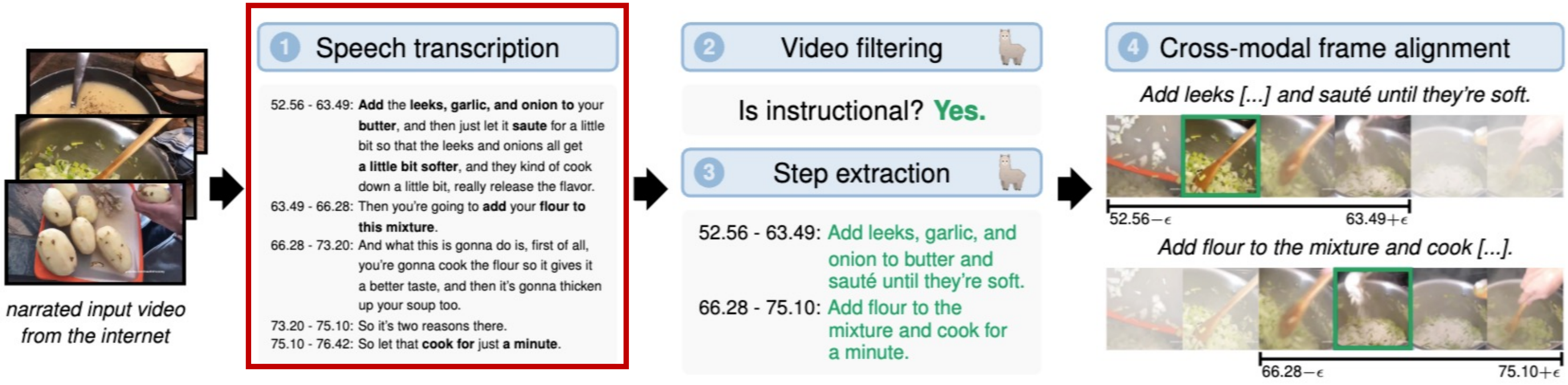
52.56 - 63.49: Add leeks, garlic, and onion to butter and sauté until they're soft.
66.28 - 75.10: Add flour to the mixture and cook for a minute.

4 Cross-modal frame alignment

Add leeks [...] and sauté until they're soft.

Add flour to the mixture and cook [...].

Building a Large-Scale ShowHowTo Dataset



narrated input video from the internet

1 Speech transcription

52.56 - 63.49: **Add** the **leeks, garlic, and onion to** your **butter**, and then just let it **saute** for a little bit so that the leeks and onions all get **a little bit softer**, and they kind of cook down a little bit, really release the flavor.
63.49 - 66.28: Then you're going to **add** your **flour to this mixture**.
66.28 - 73.20: And what this is gonna do is, first of all, you're gonna cook the flour so it gives it a better taste, and then it's gonna thicken up your soup too.
73.20 - 75.10: So it's two reasons there.
75.10 - 76.42: So let that **cook for** just **a minute**.

2 Video filtering

Is instructional? **Yes.**

3 Step extraction

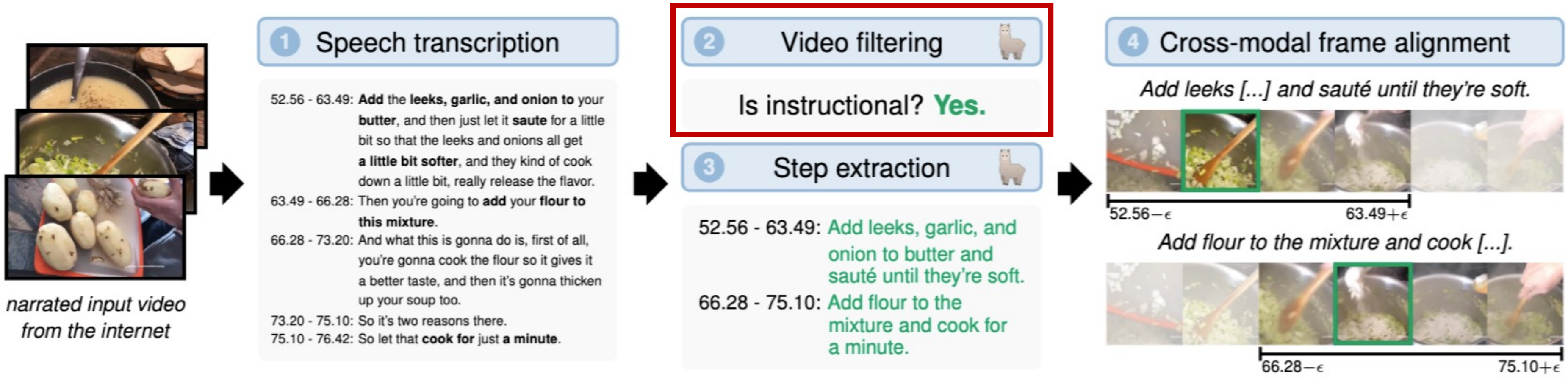
52.56 - 63.49: Add leeks, garlic, and onion to butter and sauté until they're soft.
66.28 - 75.10: Add flour to the mixture and cook for a minute.

4 Cross-modal frame alignment

Add leeks [...] and sauté until they're soft.

Add flour to the mixture and cook [...].

Building a Large-Scale ShowHowTo Dataset



narrated input video from the internet

1 Speech transcription

52.56 - 63.49: **Add** the **leeks, garlic, and onion to** your **butter**, and then just let it **saute** for a little bit so that the leeks and onions all get **a little bit softer**, and they kind of cook down a little bit, really release the flavor.

63.49 - 66.28: Then you're going to **add** your **flour to this mixture**.

66.28 - 73.20: And what this is gonna do is, first of all, you're gonna cook the flour so it gives it a better taste, and then it's gonna thicken up your soup too.

73.20 - 75.10: So it's two reasons there.

75.10 - 76.42: So let that **cook for** just **a minute**.

2 Video filtering

Is instructional? **Yes.**

3 Step extraction

52.56 - 63.49: Add leeks, garlic, and onion to butter and sauté until they're soft.

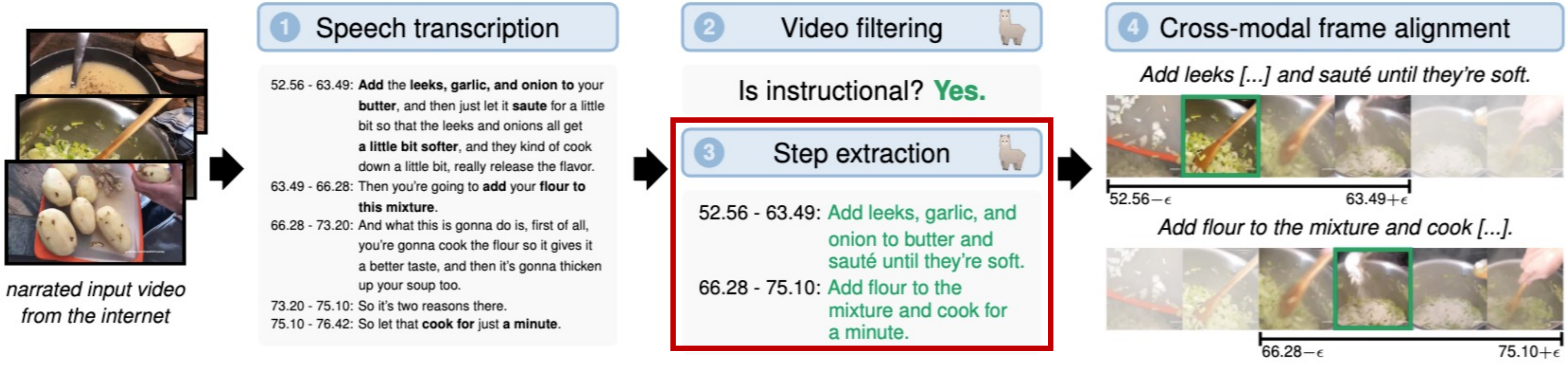
66.28 - 75.10: Add flour to the mixture and cook for a minute.

4 Cross-modal frame alignment

Add leeks [...] and sauté until they're soft.

Add flour to the mixture and cook [...].

Building a Large-Scale ShowHowTo Dataset



narrated input video from the internet

1 Speech transcription

52.56 - 63.49: **Add** the **leeks, garlic, and onion to** your **butter**, and then just let it **saute** for a little bit so that the leeks and onions all get **a little bit softer**, and they kind of cook down a little bit, really release the flavor.

63.49 - 66.28: Then you're going to **add** your **flour to this mixture**.

66.28 - 73.20: And what this is gonna do is, first of all, you're gonna cook the flour so it gives it a better taste, and then it's gonna thicken up your soup too.

73.20 - 75.10: So it's two reasons there.

75.10 - 76.42: So let that **cook for** just **a minute**.

2 Video filtering

Is instructional? **Yes.**

3 Step extraction

52.56 - 63.49: Add leeks, garlic, and onion to butter and sauté until they're soft.

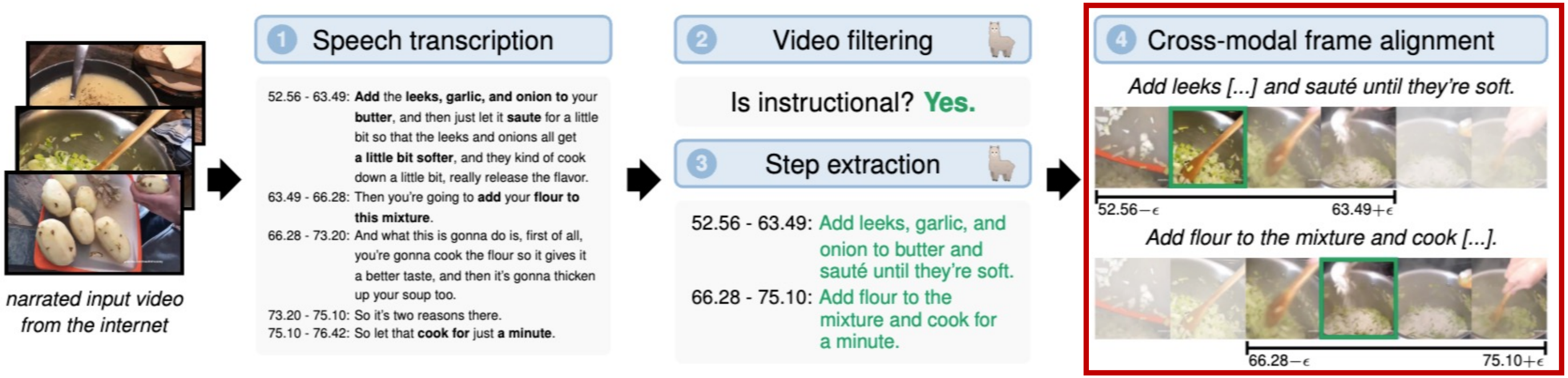
66.28 - 75.10: Add flour to the mixture and cook for a minute.

4 Cross-modal frame alignment

Add leeks [...] and sauté until they're soft.

Add flour to the mixture and cook [...].

Building a Large-Scale ShowHowTo Dataset



narrated input video from the internet

1 Speech transcription

52.56 - 63.49: **Add** the **leeks, garlic, and onion to** your **butter**, and then just let it **saute** for a little bit so that the leeks and onions all get **a little bit softer**, and they kind of cook down a little bit, really release the flavor.
63.49 - 66.28: Then you're going to **add** your **flour to this mixture**.
66.28 - 73.20: And what this is gonna do is, first of all, you're gonna cook the flour so it gives it a better taste, and then it's gonna thicken up your soup too.
73.20 - 75.10: So it's two reasons there.
75.10 - 76.42: So let that **cook for** just **a minute**.

2 Video filtering

Is instructional? **Yes.**

3 Step extraction

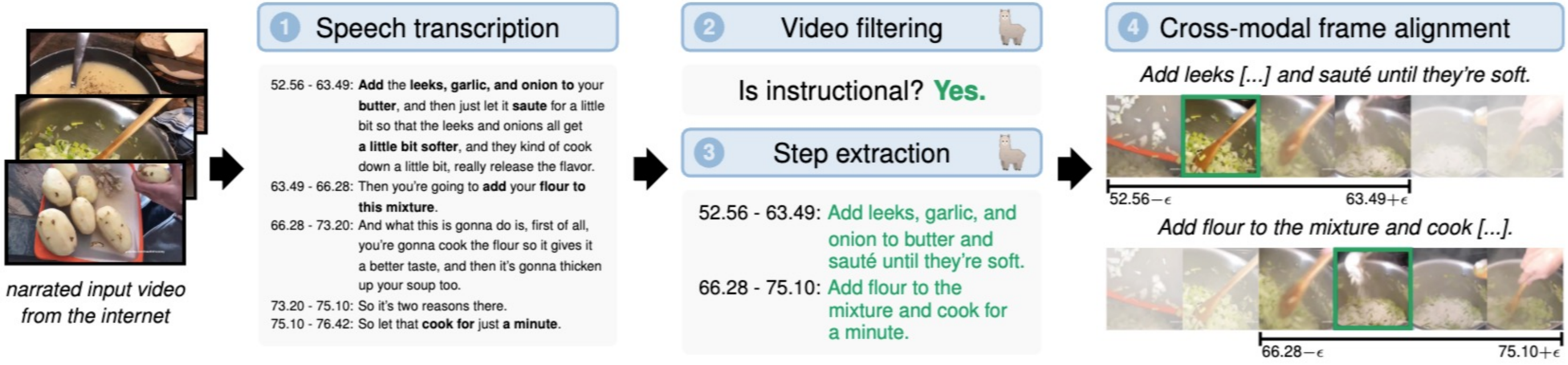
52.56 - 63.49: Add leeks, garlic, and onion to butter and sauté until they're soft.
66.28 - 75.10: Add flour to the mixture and cook for a minute.

4 Cross-modal frame alignment

Add leeks [...] and sauté until they're soft.

Add flour to the mixture and cook [...].

Building a Large-Scale ShowHowTo Dataset



narrated input video from the internet

1 Speech transcription

52.56 - 63.49: **Add** the **leeks, garlic, and onion to** your **butter**, and then just let it **saute** for a little bit so that the leeks and onions all get **a little bit softer**, and they kind of cook down a little bit, really release the flavor.
63.49 - 66.28: Then you're going to **add** your **flour to this mixture**.
66.28 - 73.20: And what this is gonna do is, first of all, you're gonna cook the flour so it gives it a better taste, and then it's gonna thicken up your soup too.
73.20 - 75.10: So it's two reasons there.
75.10 - 76.42: So let that **cook for** just **a minute**.

2 Video filtering

Is instructional? **Yes.**

3 Step extraction

52.56 - 63.49: Add leeks, garlic, and onion to butter and sauté until they're soft.
66.28 - 75.10: Add flour to the mixture and cook for a minute.

4 Cross-modal frame alignment

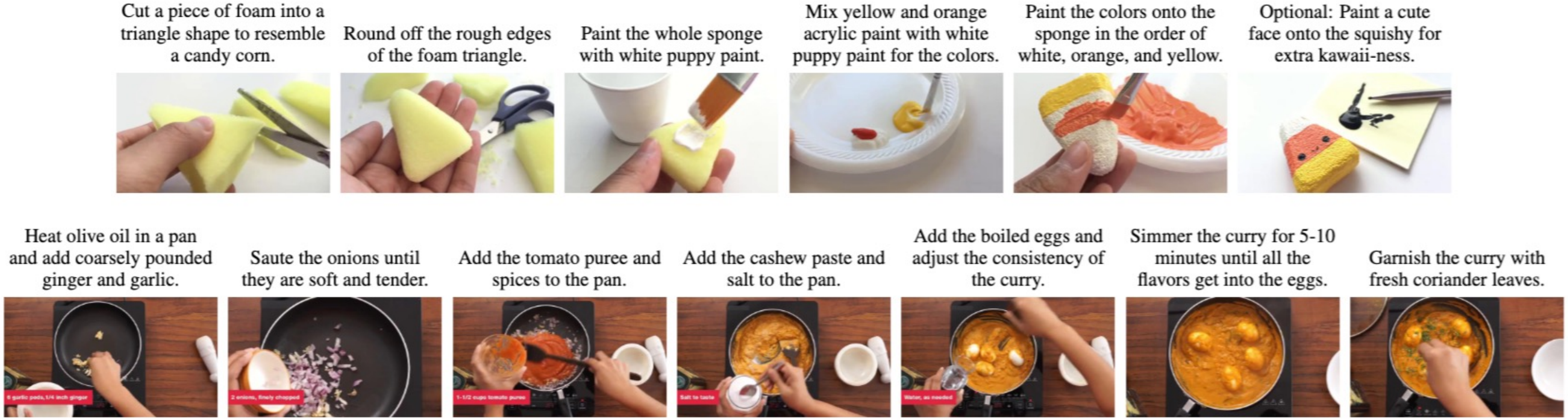
Add leeks [...] and sauté until they're soft.

Add flour to the mixture and cook [...].

Processing 1M instructional videos in HowTo100M leads to...

- A large-scale dataset (578K sequences, 4.5M steps)
- Task diversity (25K+ HowTo tasks)
- Ability to scale further (no manual annotation required)

Examples from ShowHowTo Dataset



Cut a piece of foam into a triangle shape to resemble a candy corn.

Round off the rough edges of the foam triangle.

Paint the whole sponge with white puppy paint.

Mix yellow and orange acrylic paint with white puppy paint for the colors.

Paint the colors onto the sponge in the order of white, orange, and yellow.

Optional: Paint a cute face onto the squishy for extra kawaii-ness.

Heat olive oil in a pan and add coarsely pounded ginger and garlic.

Saute the onions until they are soft and tender.

Add the tomato puree and spices to the pan.

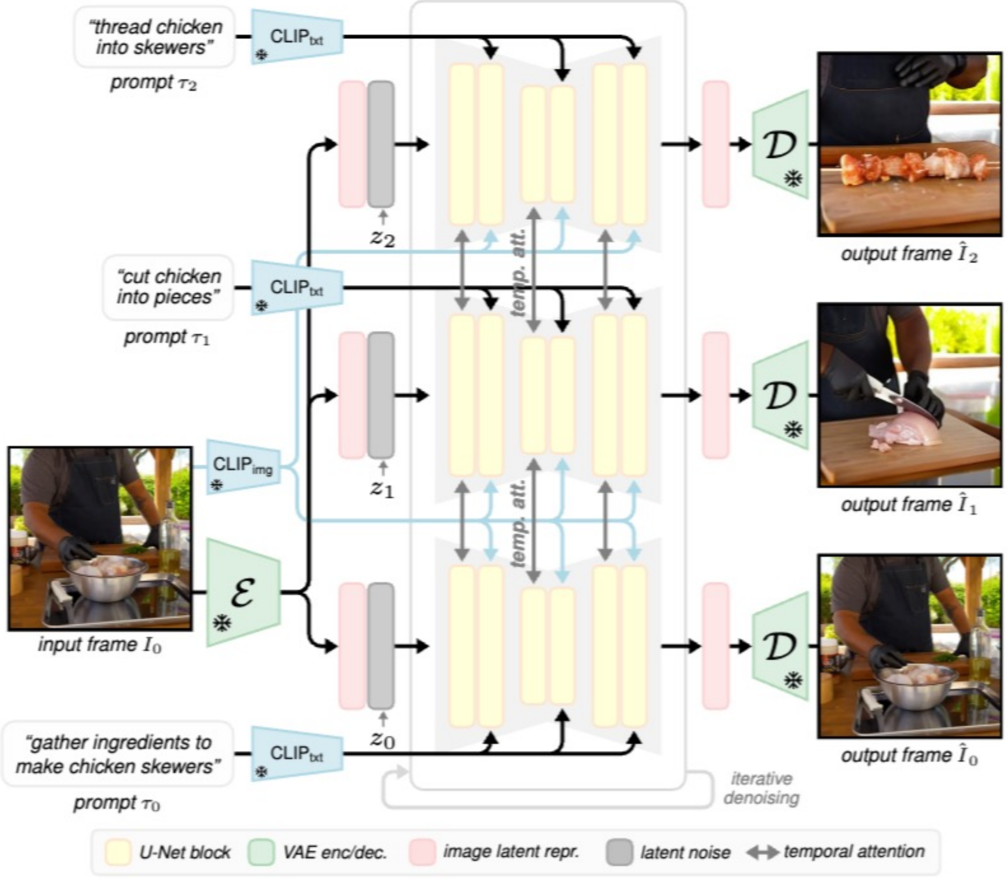
Add the cashew paste and salt to the pan.

Add the boiled eggs and adjust the consistency of the curry.

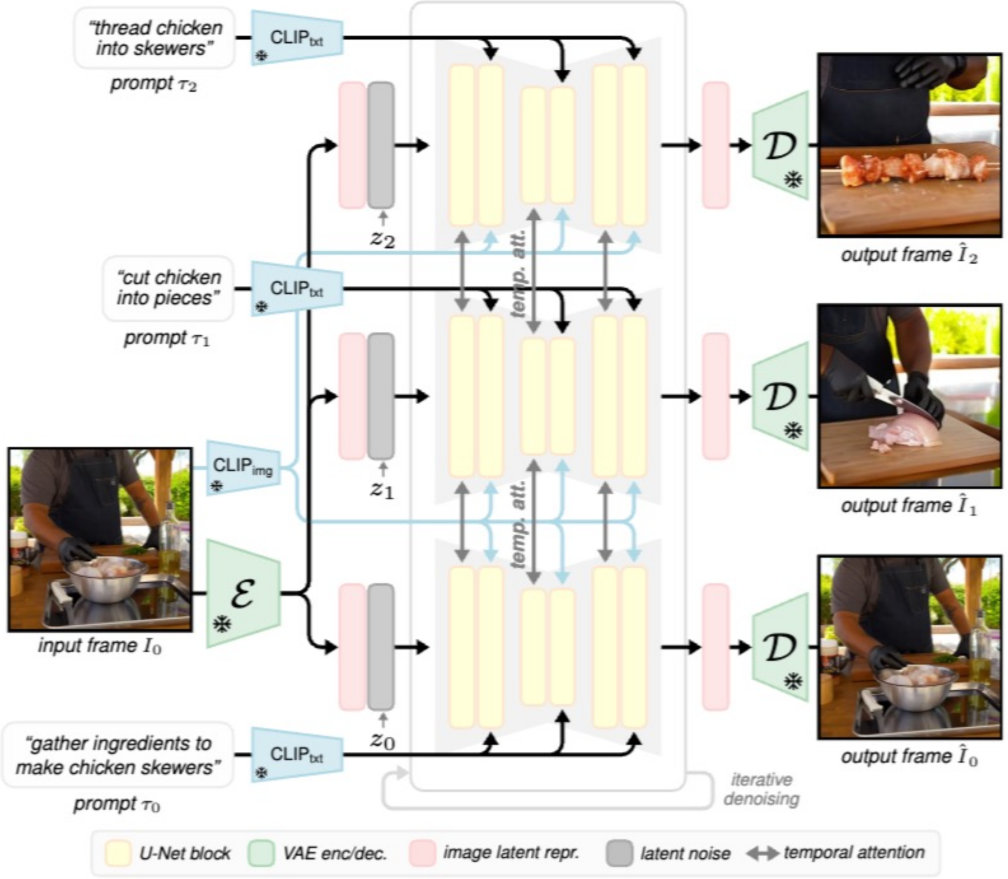
Simmer the curry for 5-10 minutes until all the flavors get into the eggs.

Garnish the curry with fresh coriander leaves.

ShowHowTo Model

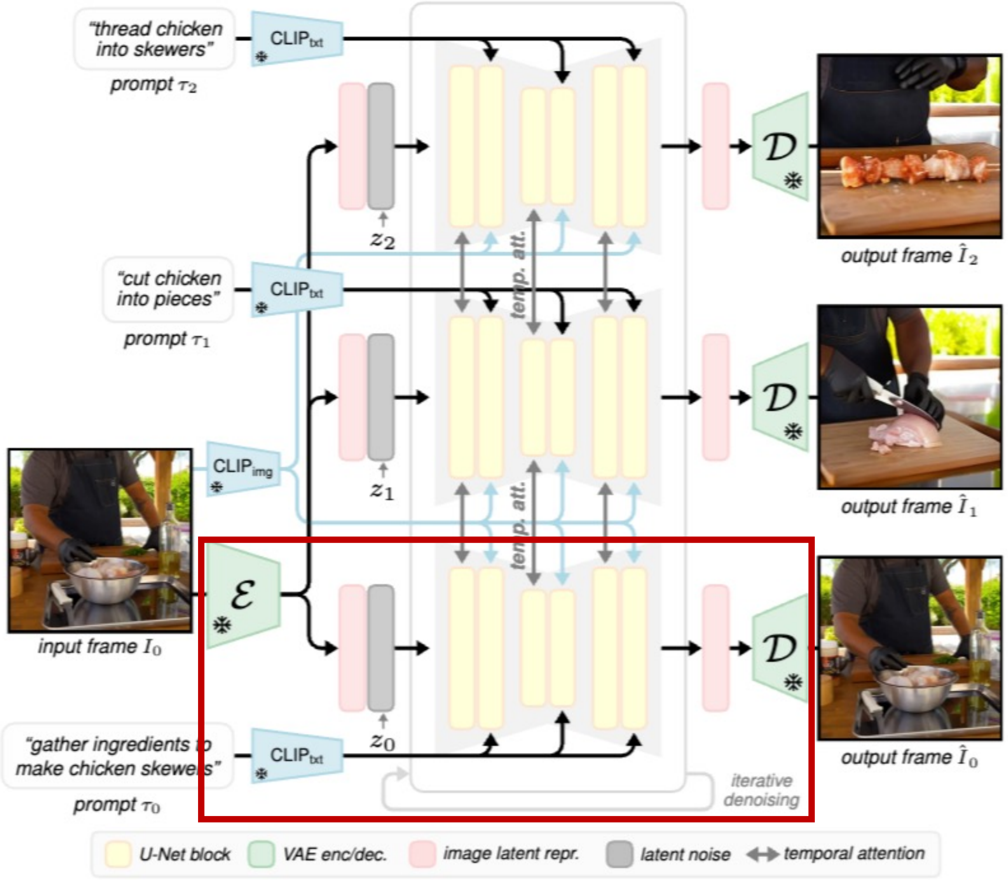


ShowHowTo Model



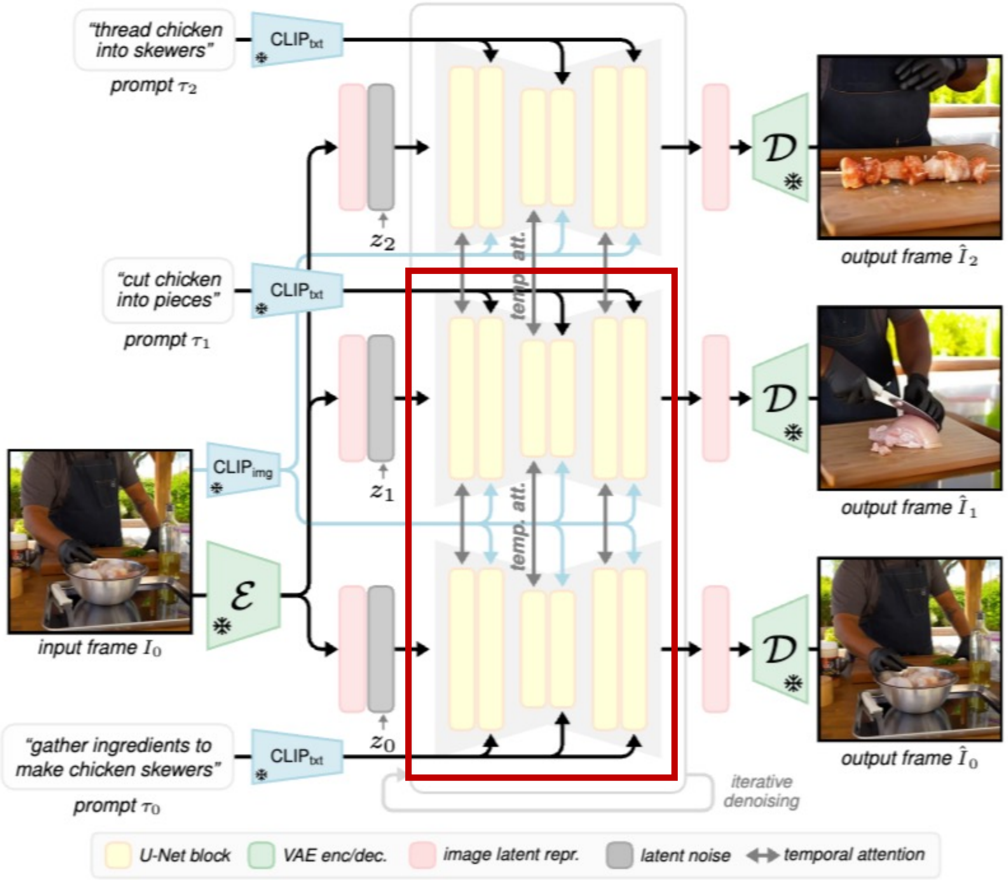
- ➢ Benefits from a pretrained video generation model (DynamiCrafter)

ShowHowTo Model



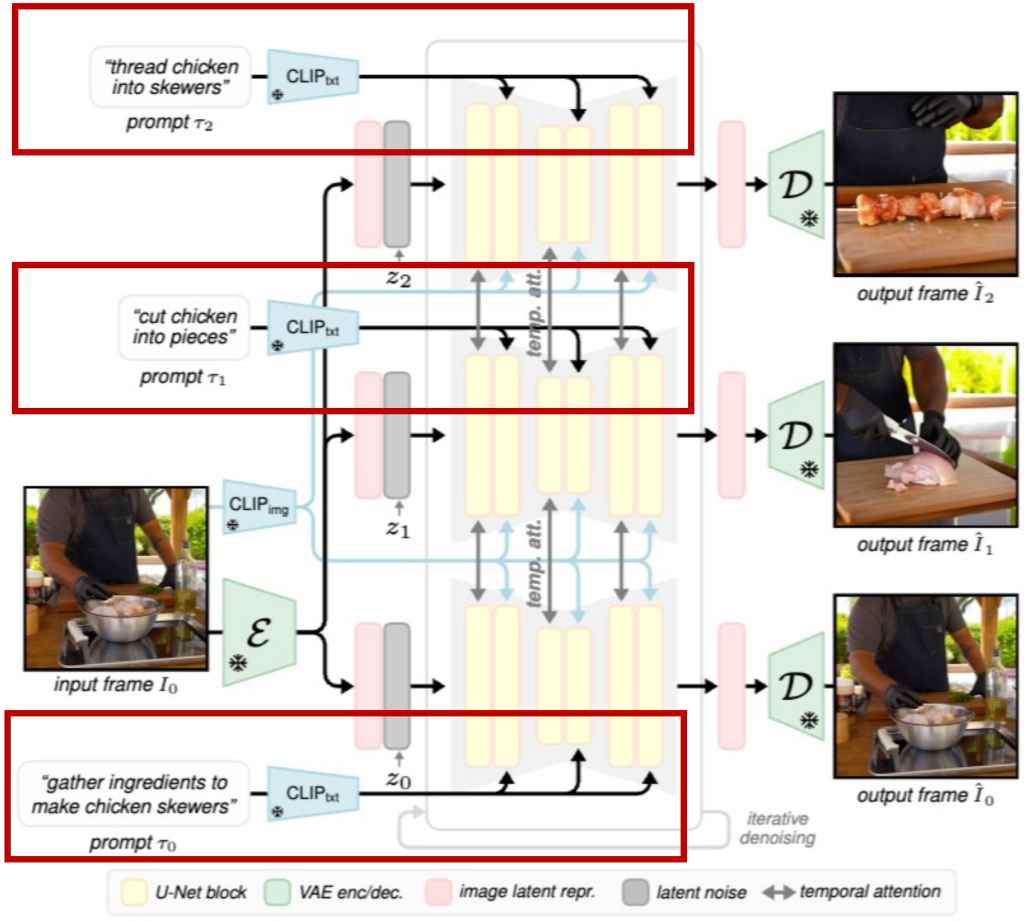
- Benefits from a pretrained video generation model (DynamiCrafter)

ShowHowTo Model



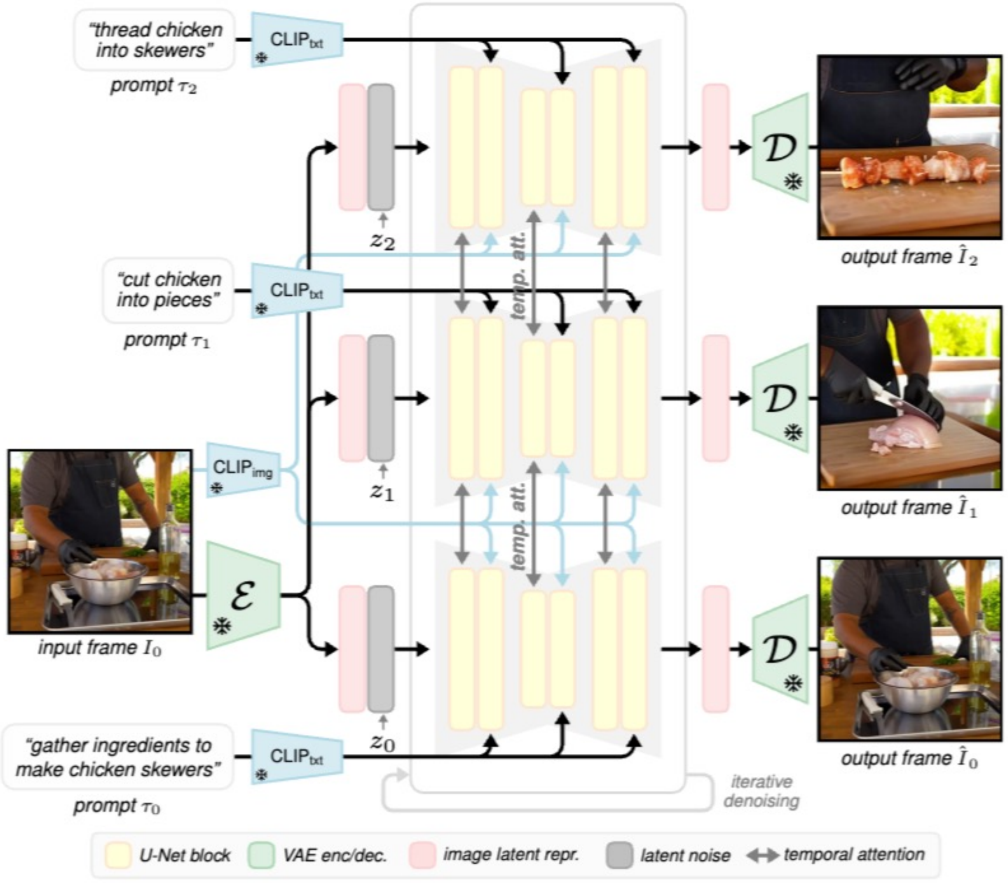
- ➢ Benefits from a pretrained video generation model (DynamiCrafter)

ShowHowTo Model



- ➢ Benefits from a pretrained video generation model (DynamiCrafter)
- ➢ Per-frame text conditioning

ShowHowTo Model



- ➢ Benefits from a pretrained video generation model (DynamiCrafter)
- ➢ Per-frame text conditioning
- ➢ Handles variable sequence-length generations

Evaluation

Knead the dough

Punch down the dough and knead

Fill the dough with ham, cheese, red sauce, mushrooms, and onions.

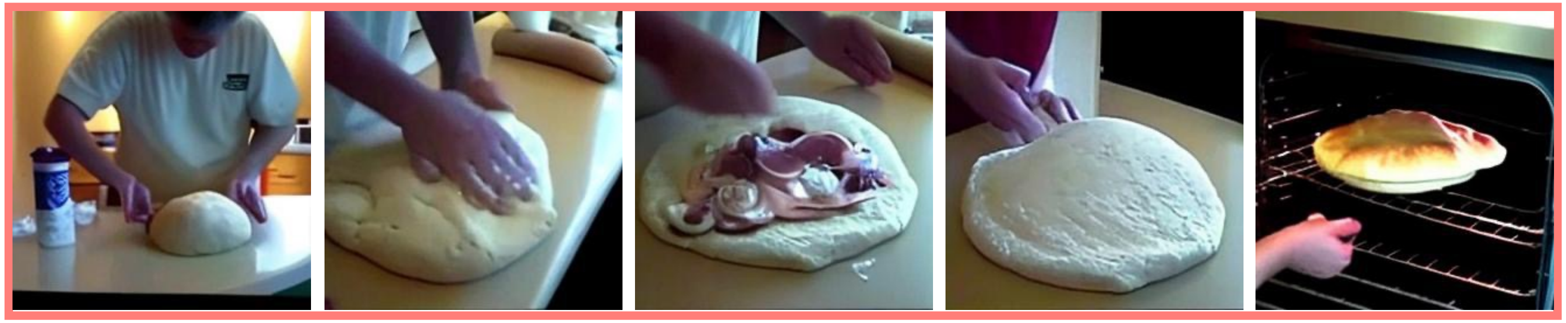
Seal the dough and let it rise again for 20-30 minutes.

Preheat the oven and bake the calzones for 6-9 minutes.

Input Image

Evaluation

Input Image



Knead the dough

Punch down the dough and knead

Fill the dough with ham, cheese, red sauce, mushrooms, and onions.

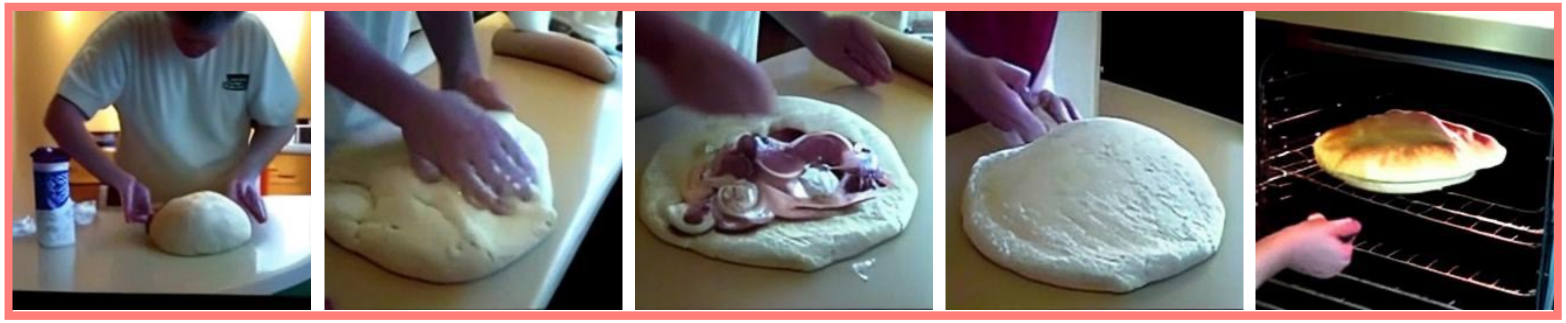
Seal the dough and let it rise again for 20-30 minutes.

Preheat the oven and bake the calzones for 6-9 minutes.

Model Generations

Evaluation

Input Image



Knead the dough

Punch down the dough and knead

Fill the dough with ham, cheese, red sauce, mushrooms, and onions.

Seal the dough and let it rise again for 20-30 minutes.

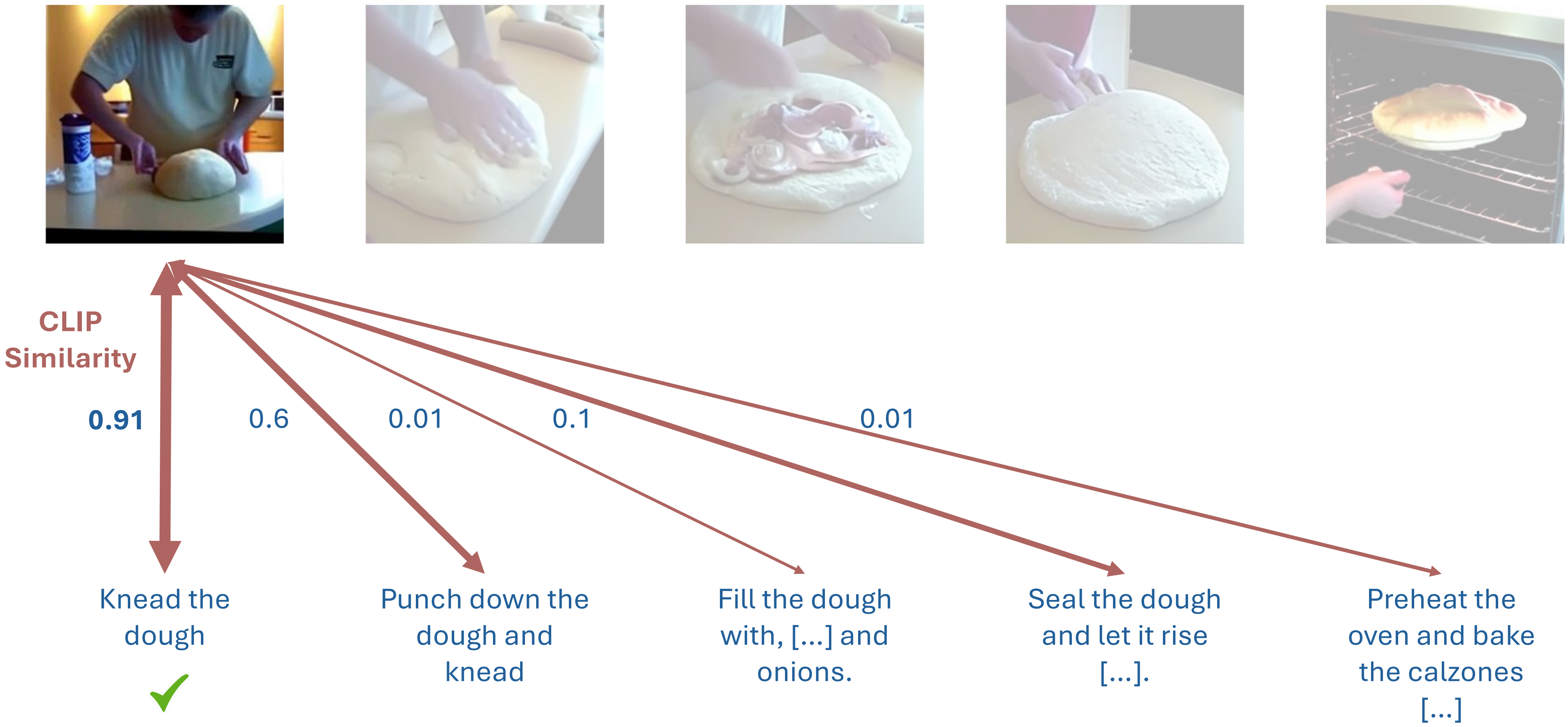
Preheat the oven and bake the calzones for 6-9 minutes.

Model Generations

But is it the correct

- Step?
- Scene?
- Task?

Step Evaluation



CLIP Similarity

0.91 0.6 0.01 0.1 0.01

Knead the dough ✓

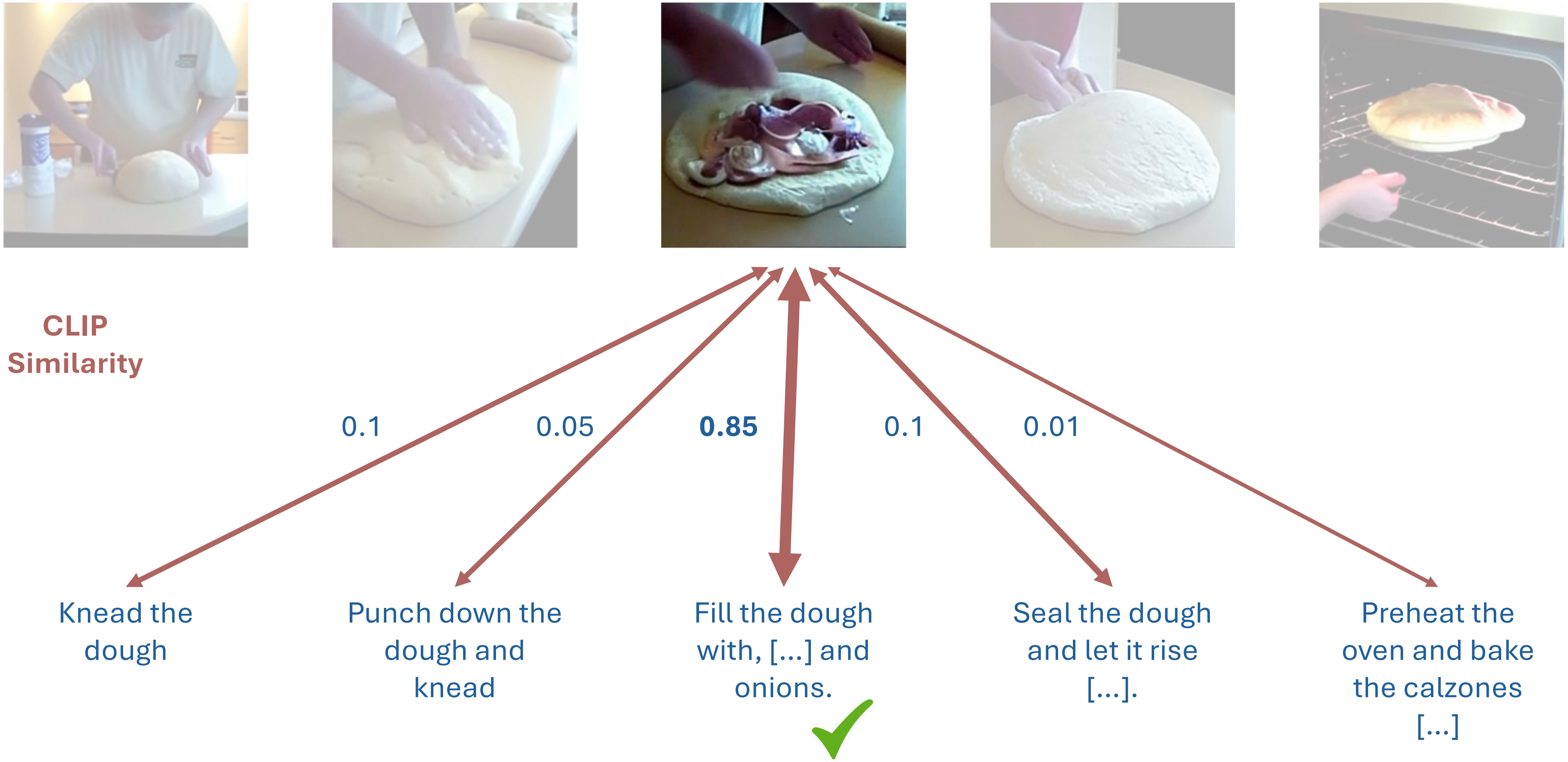
Punch down the dough and knead

Fill the dough with, [...] and onions.

Seal the dough and let it rise [...].

Preheat the oven and bake the calzones [...]

Step Faithfulness



CLIP Similarity

0.1 0.05 **0.85** 0.1 0.01

Knead the dough

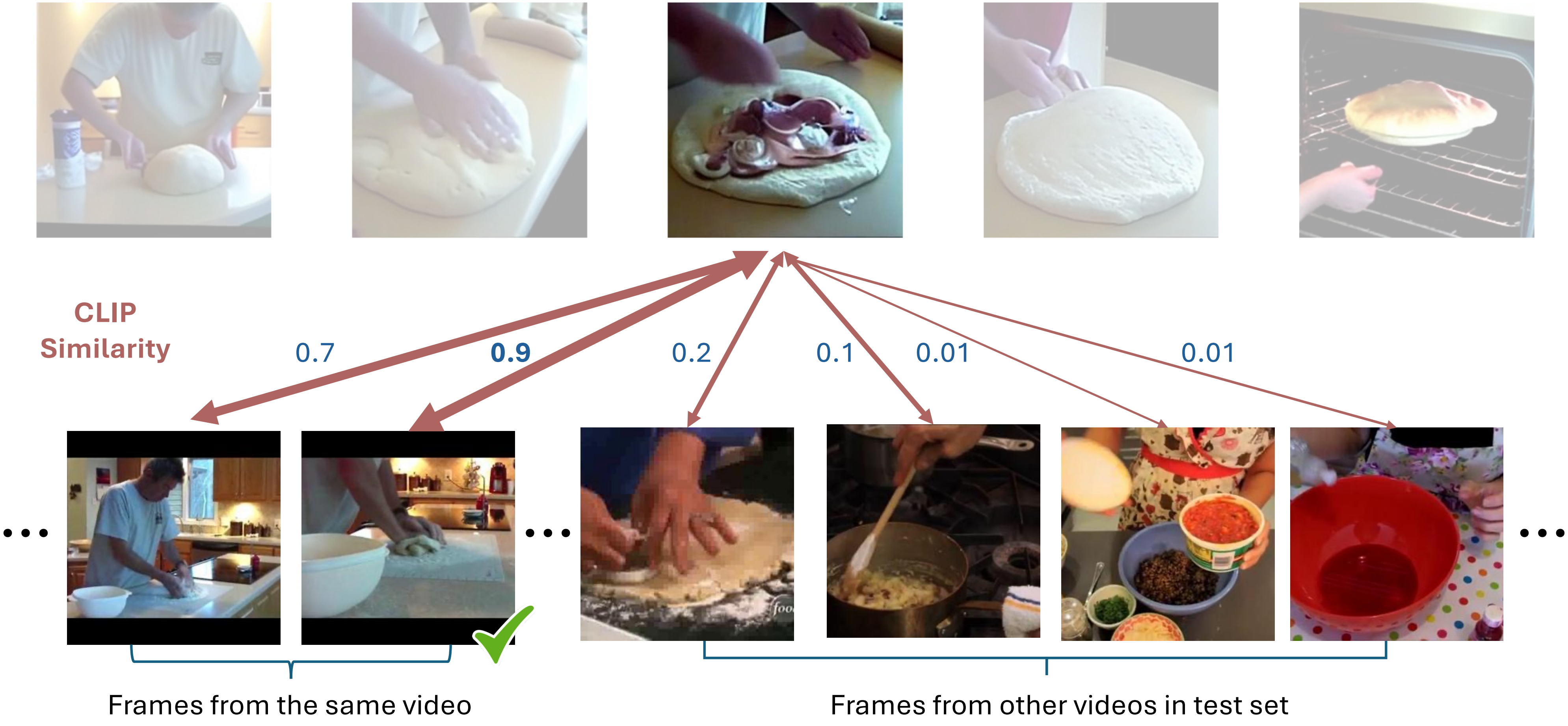
Punch down the dough and knead

Fill the dough with, [...] and onions. ✓

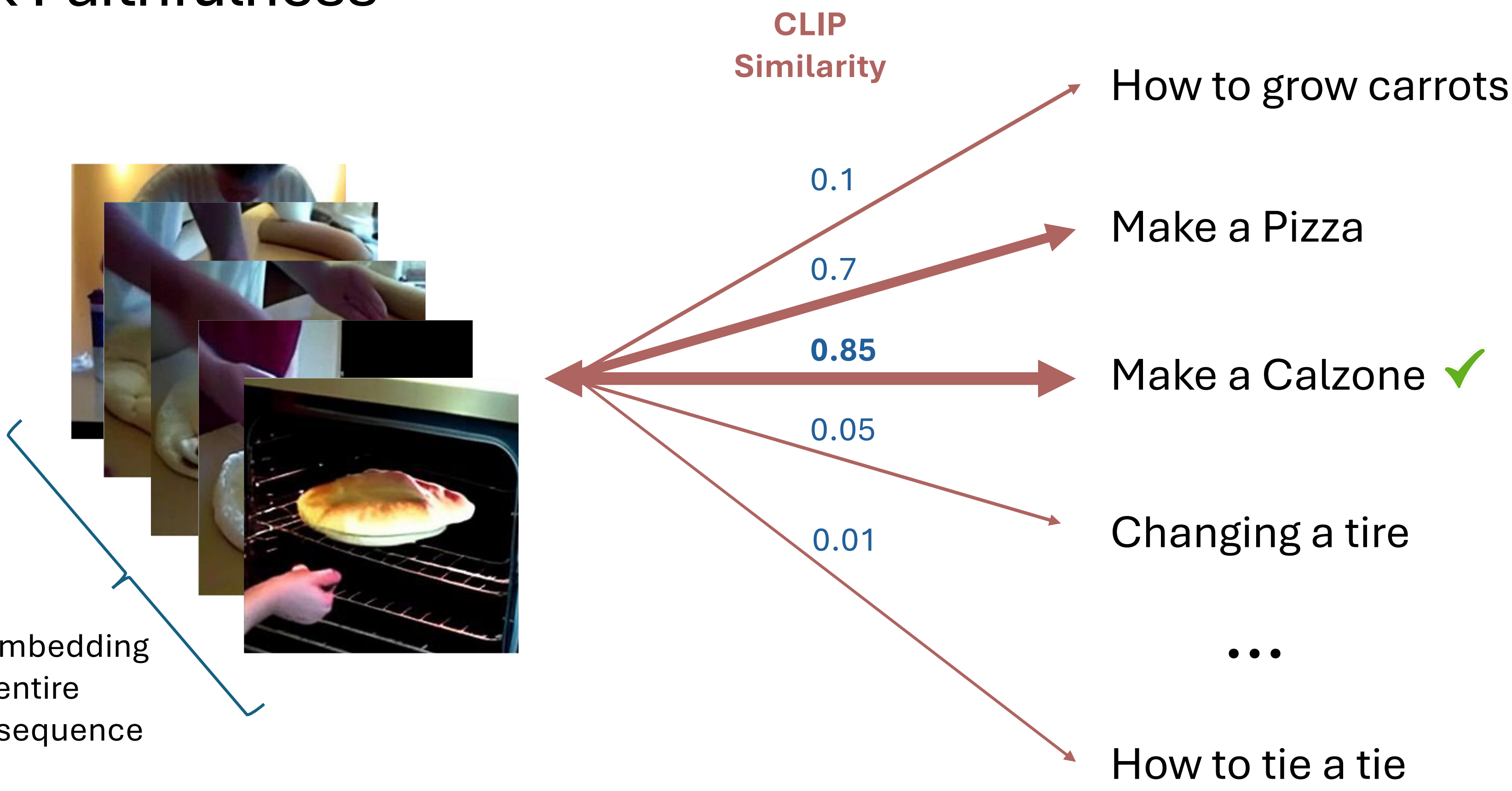
Seal the dough and let it rise [...].

Preheat the oven and bake the calzones [...]

Scene Consistency


CLIP
Similarity
0.7
0.9
0.2
0.1
0.01
0.01
Frames from the same video
Frames from other videos in test set

Task Faithfulness



CLIP Similarity

0.1

0.7

0.85

0.05

0.01

How to grow carrots

Make a Pizza

Make a Calzone ✓

Changing a tire

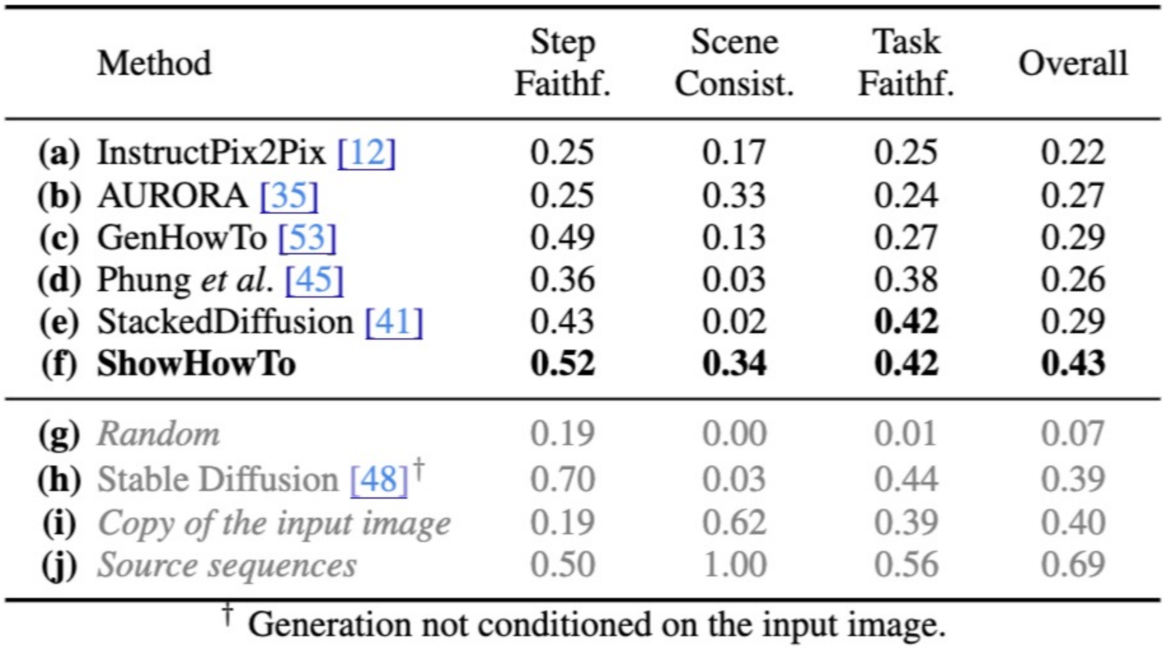
...

How to tie a tie

Avg. CLIP embedding
of the entire
generated sequence

[N=200 tasks in test set]

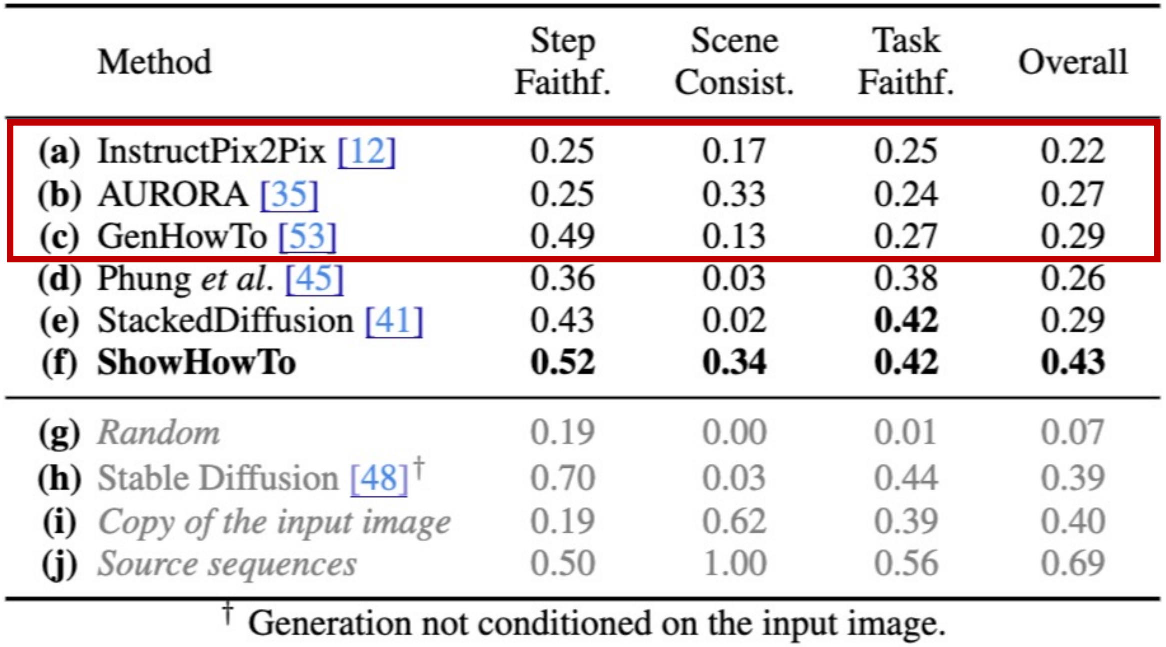
Results



	Method	Step Faithf.	Scene Consist.	Task Faithf.	Overall
(a)	InstructPix2Pix [12]	0.25	0.17	0.25	0.22
(b)	AURORA [35]	0.25	0.33	0.24	0.27
(c)	GenHowTo [53]	0.49	0.13	0.27	0.29
(d)	Phung *et al.* [45]	0.36	0.03	0.38	0.26
(e)	StackedDiffusion [41]	0.43	0.02	**0.42**	0.29
(f)	**ShowHowTo**	**0.52**	**0.34**	**0.42**	**0.43**
(g)	*Random*	0.19	0.00	0.01	0.07
(h)	Stable Diffusion [48]†	0.70	0.03	0.44	0.39
(i)	*Copy of the input image*	0.19	0.62	0.39	0.40
(j)	*Source sequences*	0.50	1.00	0.56	0.69

† Generation not conditioned on the input image.

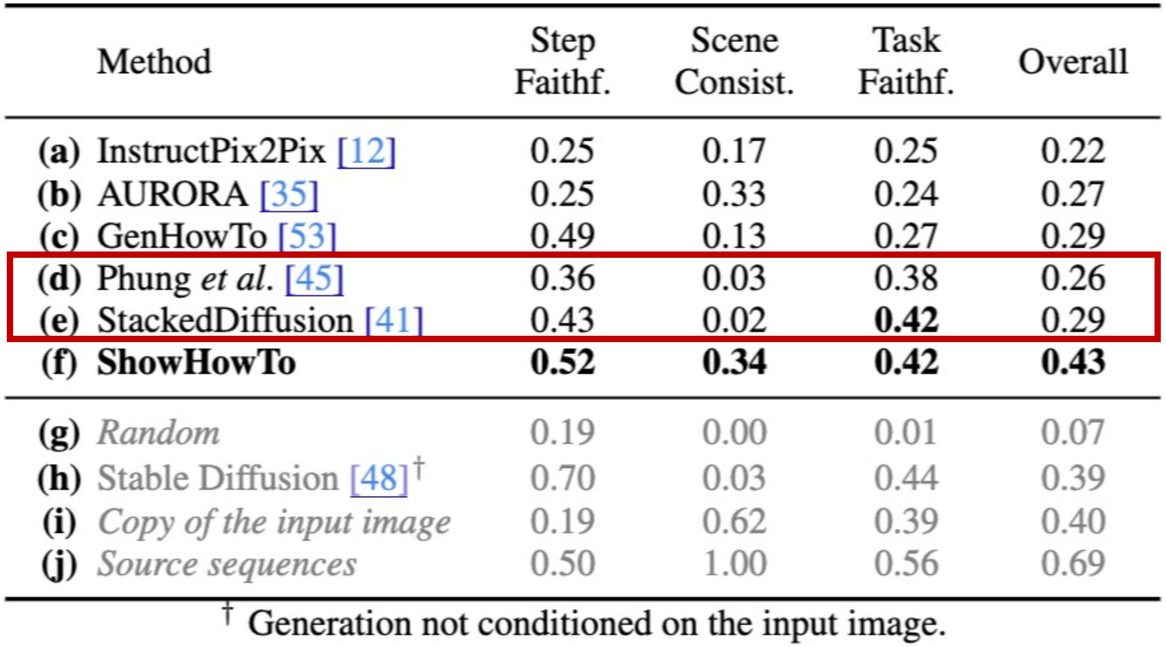
Results



	Method	Step Faithf.	Scene Consist.	Task Faithf.	Overall
(a)	InstructPix2Pix [12]	0.25	0.17	0.25	0.22
(b)	AURORA [35]	0.25	0.33	0.24	0.27
(c)	GenHowTo [53]	0.49	0.13	0.27	0.29
(d)	Phung *et al.* [45]	0.36	0.03	0.38	0.26
(e)	StackedDiffusion [41]	0.43	0.02	**0.42**	0.29
(f)	**ShowHowTo**	**0.52**	**0.34**	**0.42**	**0.43**
(g)	*Random*	0.19	0.00	0.01	0.07
(h)	Stable Diffusion [48]†	0.70	0.03	0.44	0.39
(i)	*Copy of the input image*	0.19	0.62	0.39	0.40
(j)	*Source sequences*	0.50	1.00	0.56	0.69

† Generation not conditioned on the input image.

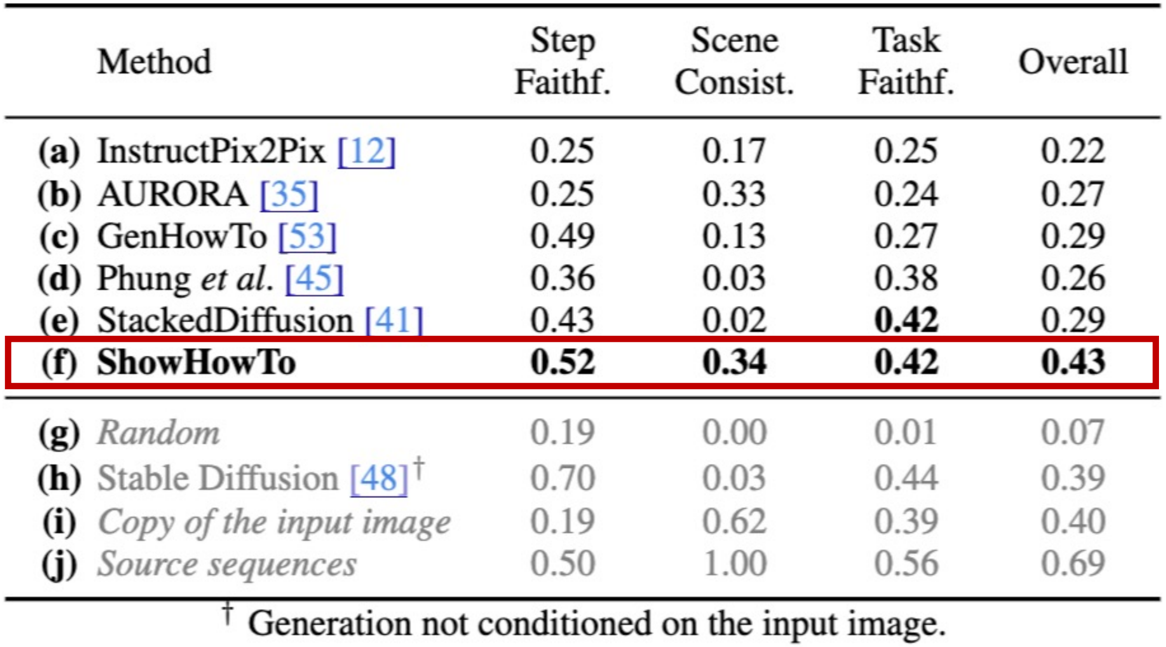
Results



	Method	Step Faithf.	Scene Consist.	Task Faithf.	Overall
(a)	InstructPix2Pix [12]	0.25	0.17	0.25	0.22
(b)	AURORA [35]	0.25	0.33	0.24	0.27
(c)	GenHowTo [53]	0.49	0.13	0.27	0.29
(d)	Phung *et al.* [45]	0.36	0.03	0.38	0.26
(e)	StackedDiffusion [41]	0.43	0.02	**0.42**	0.29
(f)	**ShowHowTo**	**0.52**	**0.34**	**0.42**	**0.43**
(g)	*Random*	0.19	0.00	0.01	0.07
(h)	Stable Diffusion [48]†	0.70	0.03	0.44	0.39
(i)	*Copy of the input image*	0.19	0.62	0.39	0.40
(j)	*Source sequences*	0.50	1.00	0.56	0.69

† Generation not conditioned on the input image.

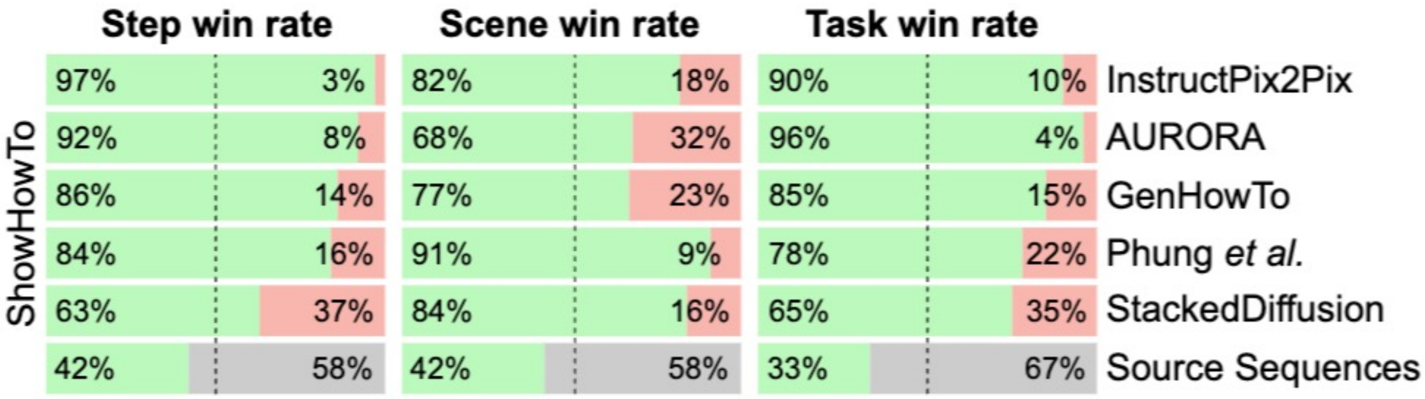
Results



	Method	Step Faithf.	Scene Consist.	Task Faithf.	Overall
(a)	InstructPix2Pix [12]	0.25	0.17	0.25	0.22
(b)	AURORA [35]	0.25	0.33	0.24	0.27
(c)	GenHowTo [53]	0.49	0.13	0.27	0.29
(d)	Phung *et al.* [45]	0.36	0.03	0.38	0.26
(e)	StackedDiffusion [41]	0.43	0.02	**0.42**	0.29
(f)	**ShowHowTo**	**0.52**	**0.34**	**0.42**	**0.43**
(g)	*Random*	0.19	0.00	0.01	0.07
(h)	Stable Diffusion [48]†	0.70	0.03	0.44	0.39
(i)	*Copy of the input image*	0.19	0.62	0.39	0.40
(j)	*Source sequences*	0.50	1.00	0.56	0.69

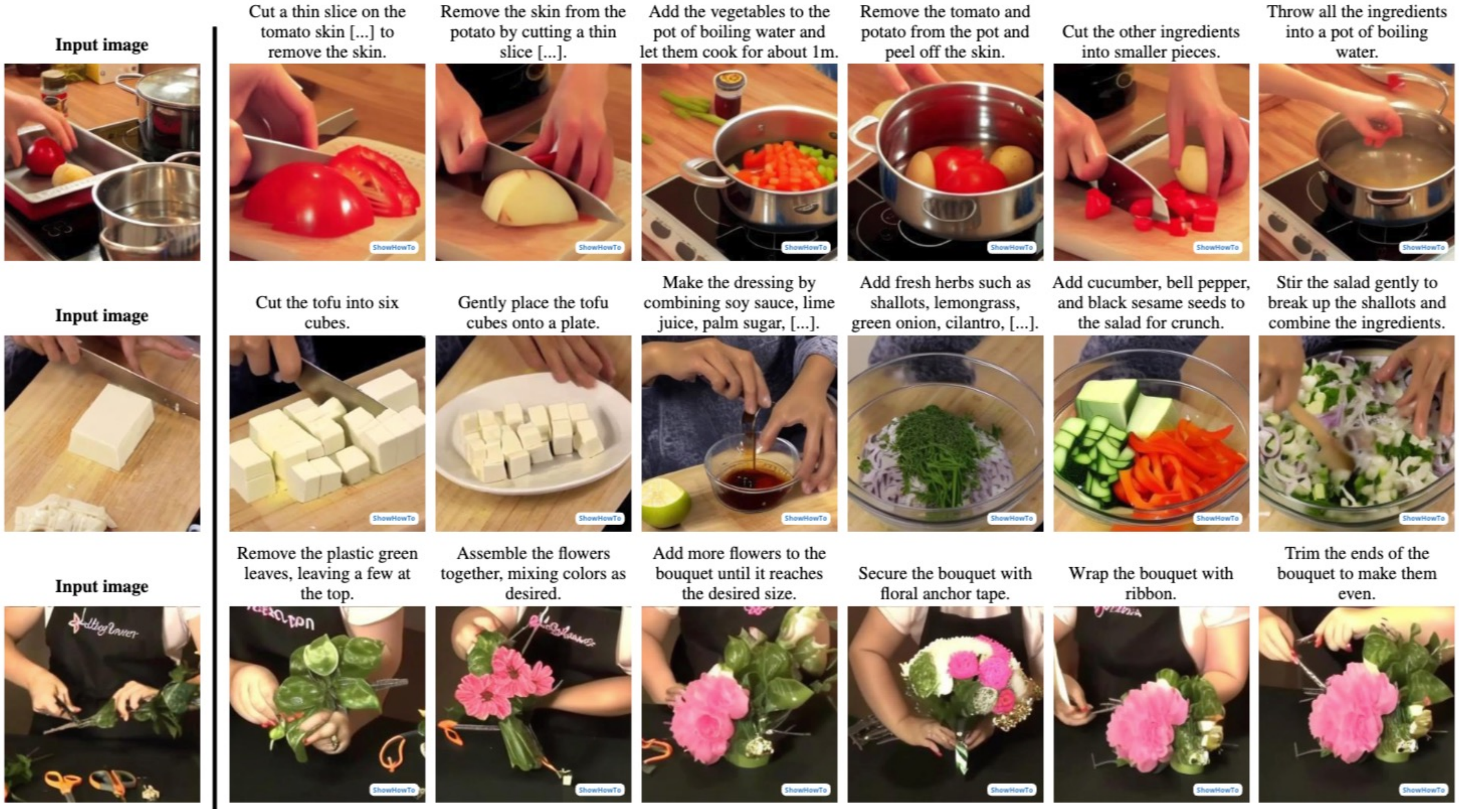
† Generation not conditioned on the input image.

User Study Results



- Participants (N=9) compared generations between ShowHowTo and baselines
- ShowHowTo was overwhelmingly preferred
- In some cases, ShowHowTo was even preferred over the source sequences

Results



Cut a thin slice on the tomato skin [...] to remove the skin.

Remove the skin from the potato by cutting a thin slice [...].

Add the vegetables to the pot of boiling water and let them cook for about 1m.

Remove the tomato and potato from the pot and peel off the skin.

Cut the other ingredients into smaller pieces.

Throw all the ingredients into a pot of boiling water.

Cut the tofu into six cubes.

Gently place the tofu cubes onto a plate.

Make the dressing by combining soy sauce, lime juice, palm sugar, [...].

Add fresh herbs such as shallots, lemongrass, green onion, cilantro, [...].

Add cucumber, bell pepper, and black sesame seeds to the salad for crunch.

Stir the salad gently to break up the shallots and combine the ingredients.

Remove the plastic green leaves, leaving a few at the top.

Assemble the flowers together, mixing colors as desired.

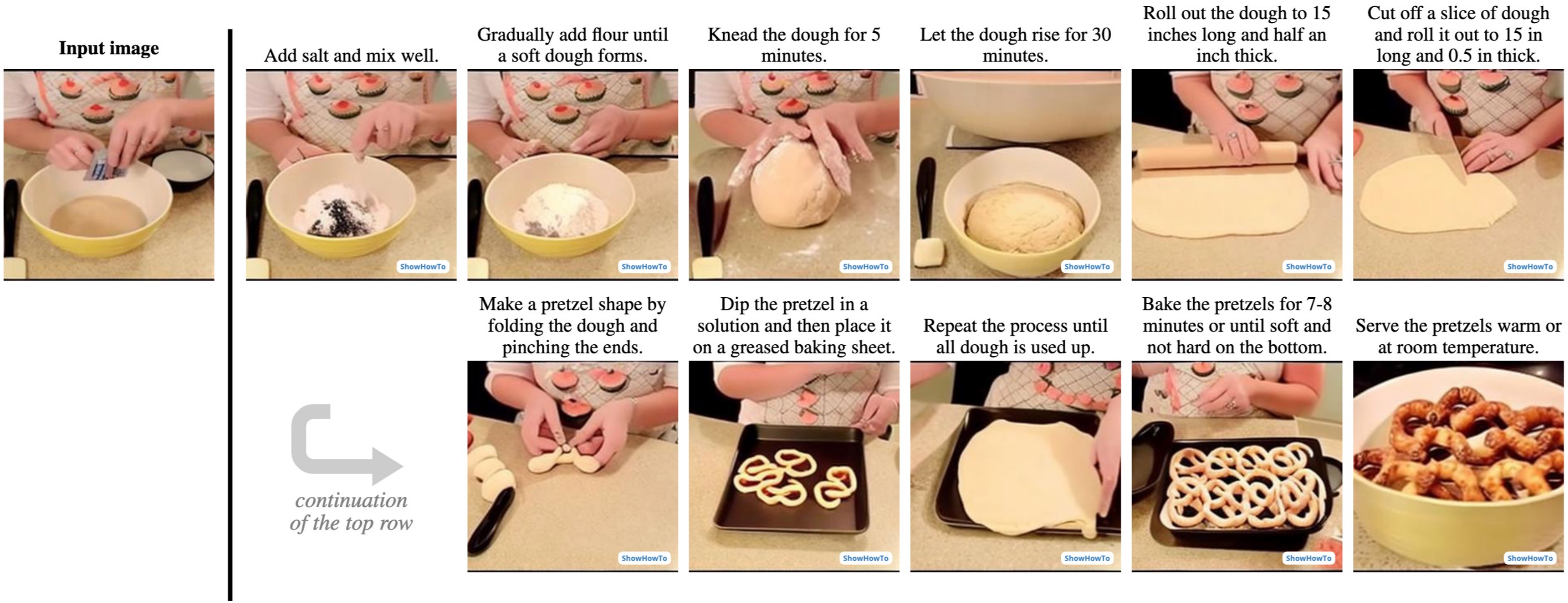
Add more flowers to the bouquet until it reaches the desired size.

Secure the bouquet with floral anchor tape.

Wrap the bouquet with ribbon.

Trim the ends of the bouquet to make them even.

Long sequence generation



Input image

Add salt and mix well.

Gradually add flour until a soft dough forms.

Knead the dough for 5 minutes.

Let the dough rise for 30 minutes.

Roll out the dough to 15 inches long and half an inch thick.

Cut off a slice of dough and roll it out to 15 in long and 0.5 in thick.

continuation of the top row

Make a pretzel shape by folding the dough and pinching the ends.

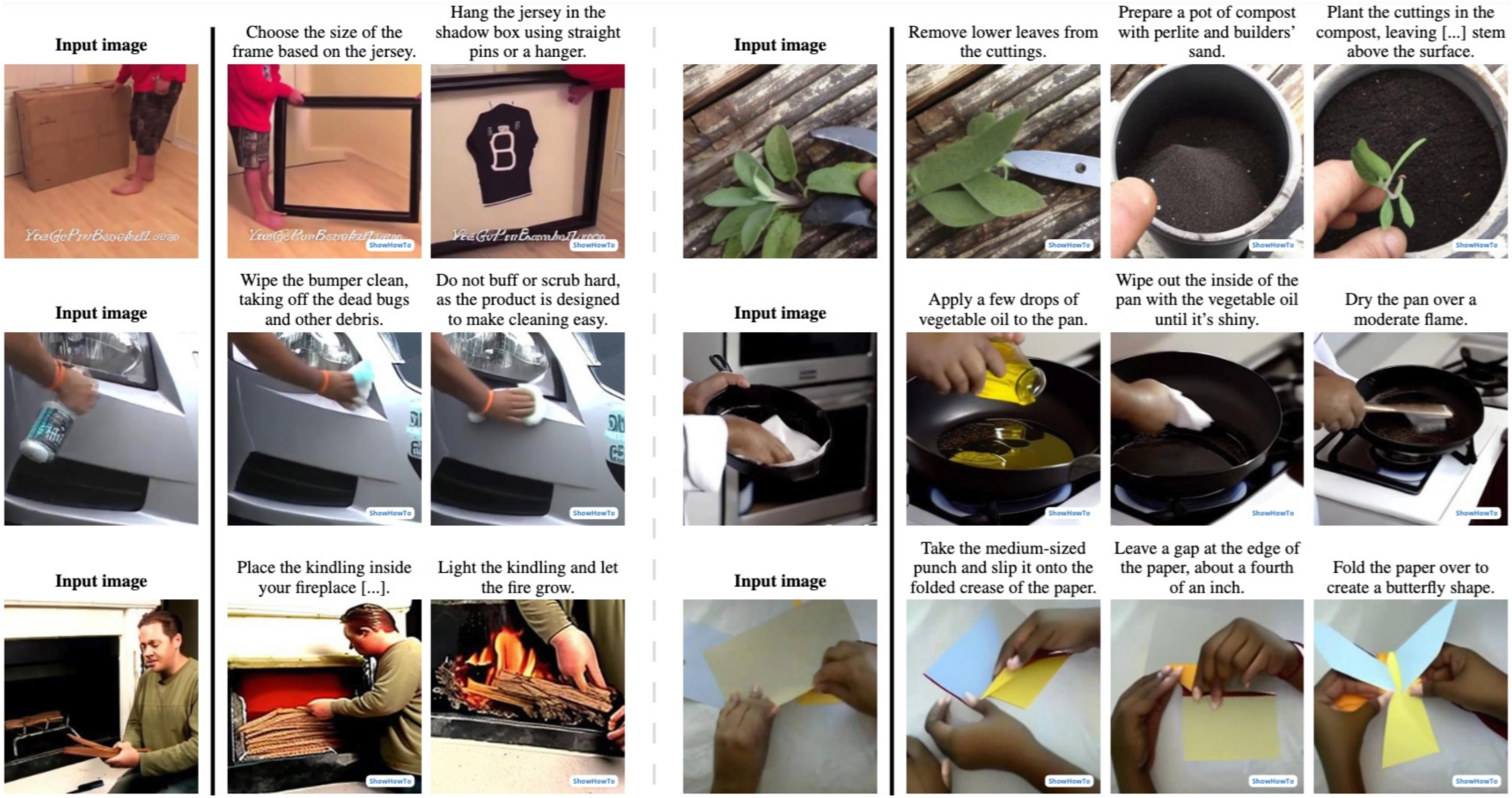
Dip the pretzel in a solution and then place it on a greased baking sheet.

Repeat the process until all dough is used up.

Bake the pretzels for 7-8 minutes or until soft and not hard on the bottom.

Serve the pretzels warm or at room temperature.

Results



Input image

Choose the size of the frame based on the jersey.

Hang the jersey in the shadow box using straight pins or a hanger.

Input image

Remove lower leaves from the cuttings.

Prepare a pot of compost with perlite and builders' sand.

Plant the cuttings in the compost, leaving [...] stem above the surface.

Input image

Wipe the bumper clean, taking off the dead bugs and other debris.

Do not buff or scrub hard, as the product is designed to make cleaning easy.

Input image

Apply a few drops of vegetable oil to the pan.

Wipe out the inside of the pan with the vegetable oil until it's shiny.

Dry the pan over a moderate flame.

Input image

Place the kindling inside your fireplace [...].

Light the kindling and let the fire grow.

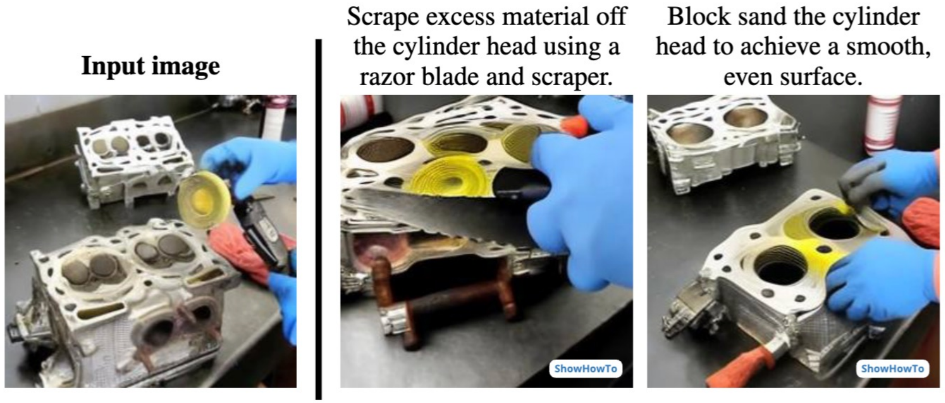
Input image

Take the medium-sized punch and slip it onto the folded crease of the paper.

Leave a gap at the edge of the paper, about a fourth of an inch.

Fold the paper over to create a butterfly shape.

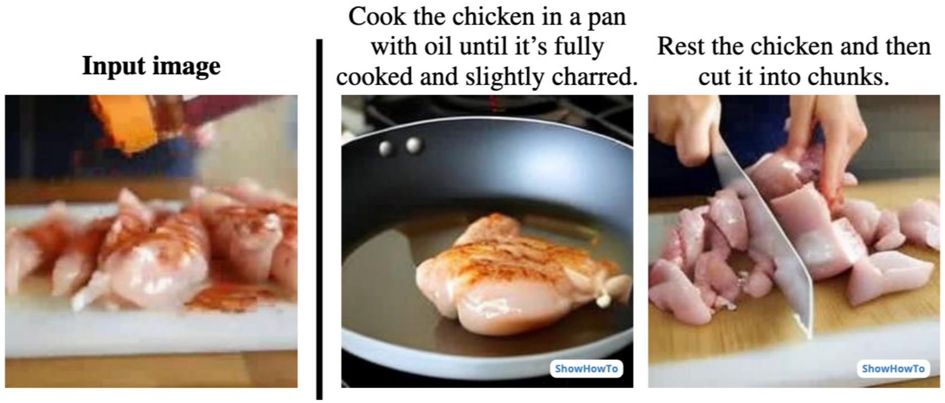
ShowHowTo model is not without limitations…



Scrape excess material off the cylinder head using a razor blade and scraper.

Block sand the cylinder head to achieve a smooth, even surface.

Can struggle with rare objects or tools



Cook the chicken in a pan with oil until it’s fully cooked and slightly charred.

Rest the chicken and then cut it into chunks.

Can fail to update object states
E.g., Raw → Cooked → Raw

Summary

- Introduced the problem of generating scene-conditioned visual instructions
- A fully automatic approach for collecting visual instruction dataset from web instructional videos
- A large-scale dataset of 0.6M visual instruction sequences, with 4.5M steps, covering 25K HowTo tasks
- Trained a video diffusion model capable of generating step-by-step visual instructions

ShowHowTo paper, code, and dataset available :-)

https://soczech.github.io/showhowto/