NOVA SCHOOL OF SCIENCE & TECHNOLOGY

NOVALINCS LABORATORY FOR COMPUTER SCIENCE AND INFORMATICS

Google Research

Generating Coherent Sequences of Visual Illustrations for Real-World Manual Tasks

João Bordalo, Vasco Ramos, Rodrigo Valério, Diogo Glória-Silva, Yonatan Bitton, Michal Yarom, Idan Szpektor, Joao Magalhaes

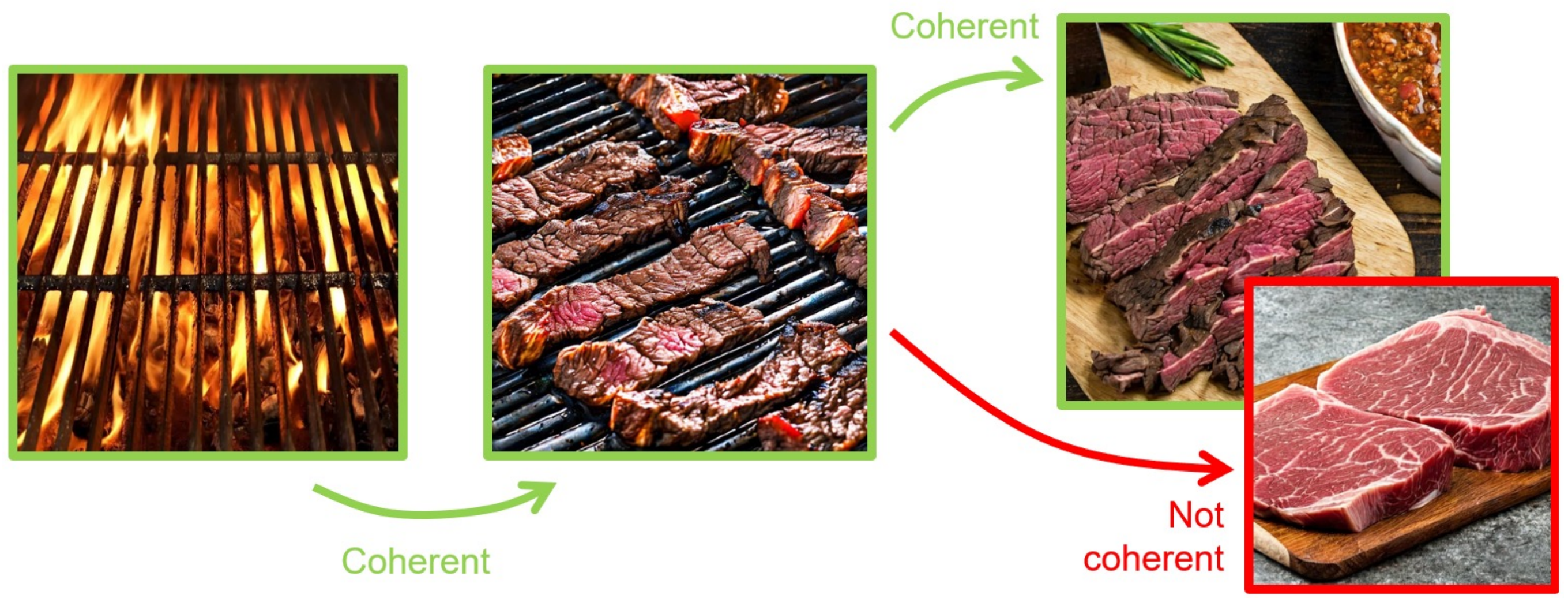
Challenges

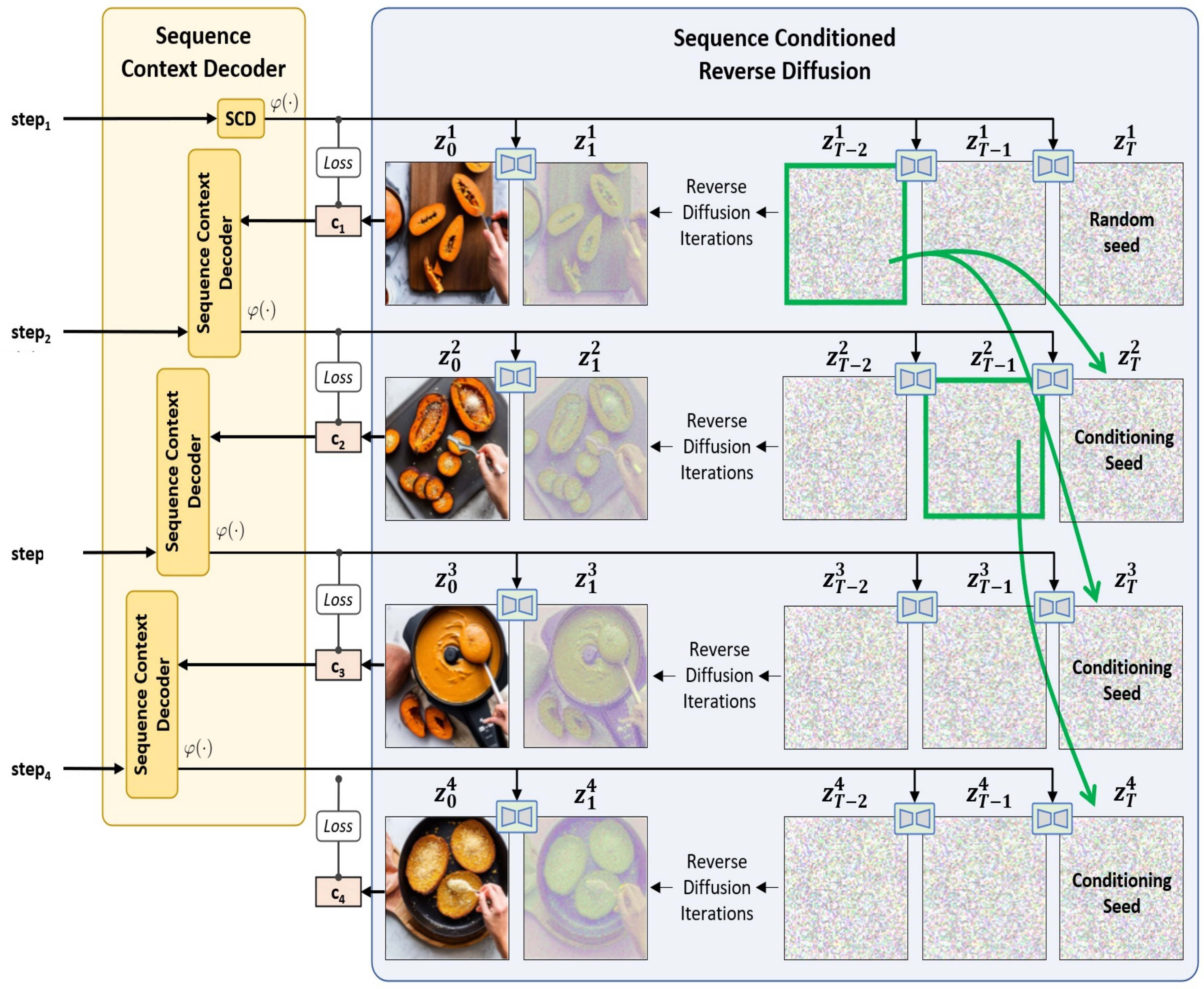
- **Clearly illustrate the actions** described in the instructions.
- Maintain **semantic coherence** by ensuring objects remain consistent across consecutive images.
- Ensure **visual coherence** with consistent backgrounds and visual properties in all images.
- **Non-linear sequence** where steps may not always relate directly to the **previous step**

Step 1
Heat the grill.

Step 2
Cook beef for about 5 minutes, each side.

Step 3
Let the beef rest on a cutting board.





Sequential Latent Diffusion Model

- Transform the step's context into a **visual caption**.
- Ensure generated captions are **contextually relevant** by considering the target step and previous steps window.
- Condition the current image generation on **previous ones**.
- Identify the **most similar previous step** to use as a base for generating the new image.
- Select the **best latent representations** from the chosen image to serve as the seed for the new generation.

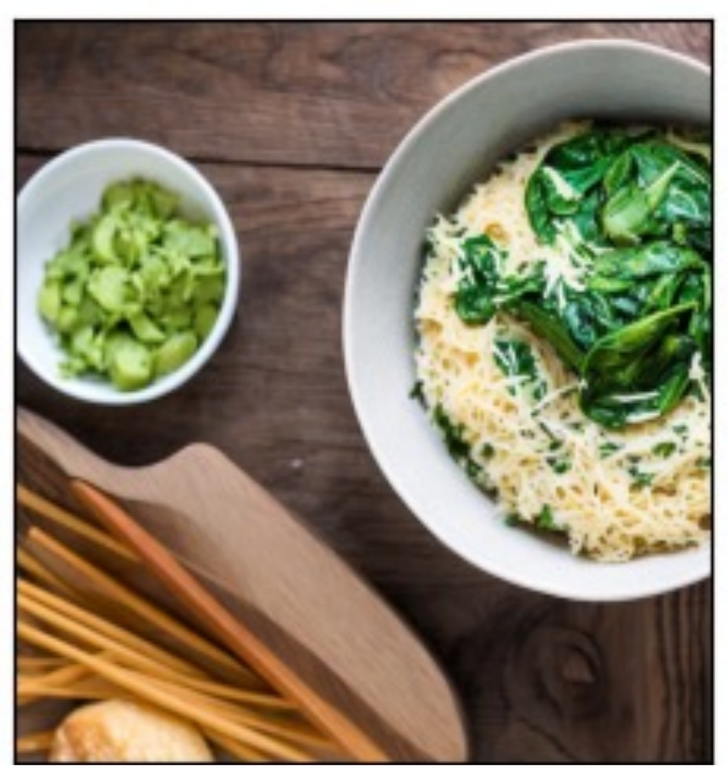
Heat the Coconut Oil in a wide pan over a medium flame, then add the Onion, Garlic, Scallion, and Ground Black Pepper. Reduce the heat to low for about 3-4 minutes.

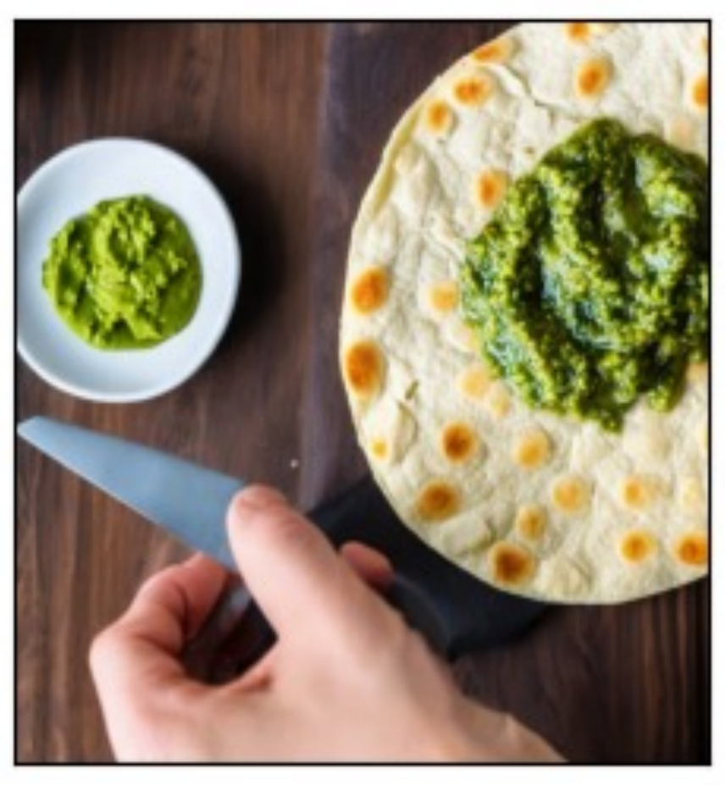
Add the Small Shrimp, stir well and cook for another 3 minutes.

Turn the heat up to medium high and add the Jamaican Callaloo, Tomato, Scotch Bonnet Pepper, Fresh Thyme, and Sea Salt. After a couple minutes, add the Water and cook until tender.

After about 10-12 minutes, taste for salt and adjust accordingly.

To a medium bowl, add Pecans, Vegan Cream Cheese, Daiya® Mozzarella Cheese, Sun-Dried Tomatoes in Olive Oil, Garlic, Ground Black Pepper, and Vegan Basil Pesto. Stir until well combined.

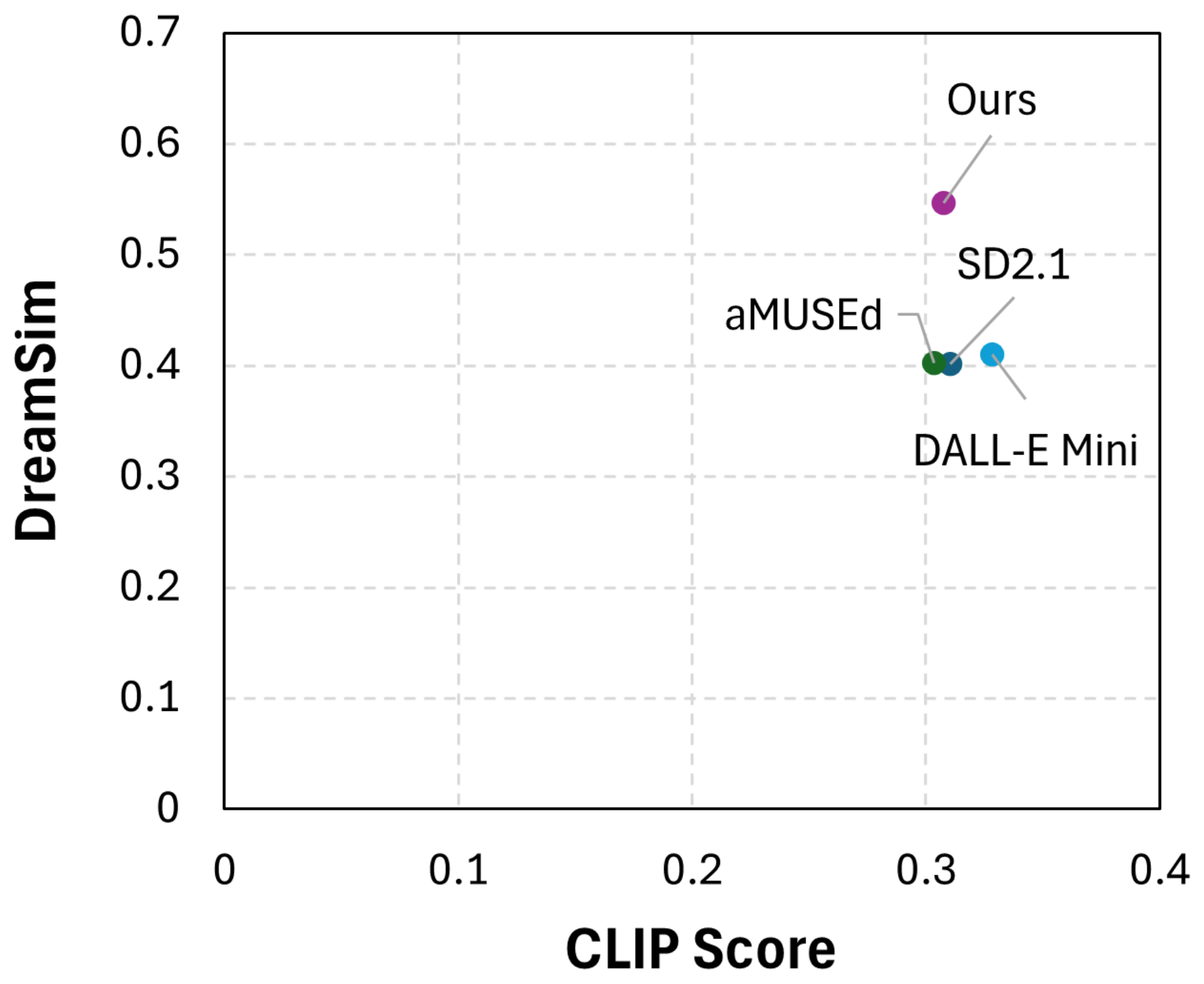

Add Fresh Spinach to the mixture and continue stirring until all ingredients are well mixed.


Lay out Tortilla, divide cheese/pesto/tomato mixture evenly among the four tortillas, spreading in a thin layer until it coats the whole tortilla up to the outer edge. Tear and sprinkle approximately 5 Fresh Basil Leaf over each tortilla.

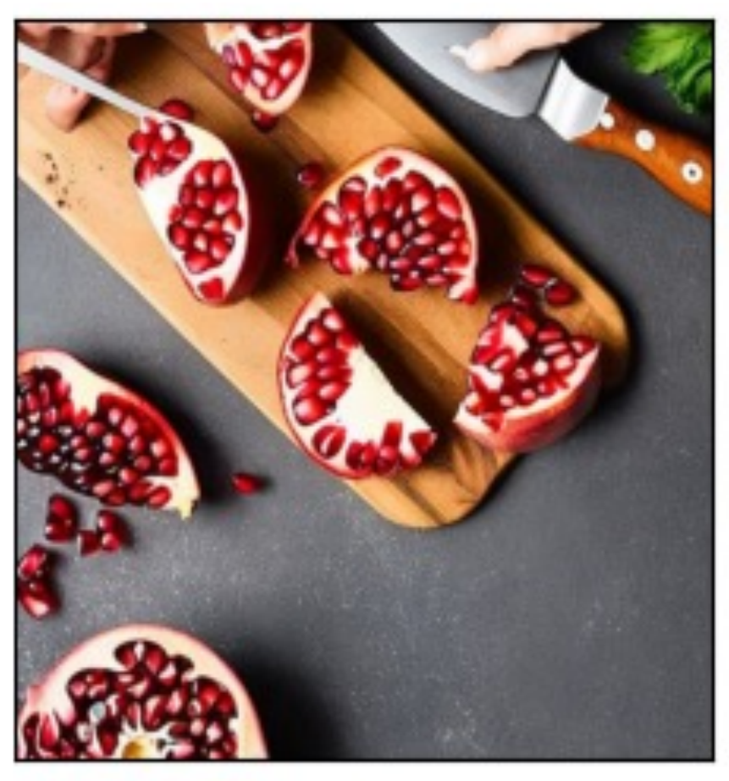
Gently roll tortillas, keeping them fairly tight, then slice each into 8 pieces. Serve immediately or store in the efrigerator for up to 3 days.

Results and Conclusions

- Improves **image sequence coherence** (DreamSim) while maintaining **text-to-image generation quality** (CLIP Score).
- **Preserves key visual and semantic traits** from selected images.
- Preferred by human annotators in both recipe and out-of-domain (DIY) tasks, ensuring better **overall sequence coherence** and **user preference**.
- Highlights the importance of selectively **conditioning the denoising process** on previous steps.



Method	Recipes (seen)	DIY (unseen)
Proposed method (wins)	**46.67**	**30.00**
Second best (wins)	26.67	23.33
Tie	10.00	16.67
No good sequence	16.67	30.00


Halve pomegranate, deseed, and set aside. Finely chop garlic and set aside. Finely chop mint. Add yogurt to a bowl and mix with chopped mint and cumin. Add salt and sugar to taste.

Preheat oven to 180°C/350°F. Heat olive oil in a pan over medium-high heat. Season lamb with salt and pepper and sear for approx. 2 min. on both sides along with chopped garlic. Remove from pan and

Transfer pita bread to oven and toast for approx. 2 min. on 180°C/350°F. Thinly slice lamb fillet.

Remove pita bread from oven and fill with yogurt sauce, pomegranate seeds, toasted pistachios, and thinly sliced lamb. Serve immediately.

This work was partially supported by Amazon Alexa Prize TaskBot, by a Google Research Gift and by NOVA LINCS ref. UIDB/04516/2020 (https://doi.org/10.54499/UIDB/04516/ 2020).