

Contrastive Sequential-Diffusion Learning: Non-linear and Multi-Scene Instructional Video Synthesis — Supplementary Material —

Vasco Ramos¹, Yonatan Bitton², Michal Yarom², Idan Szpektor², Joao Magalhaes¹

¹NOVA LINCS, NOVA School of Science and Technology, Portugal

²Google Research

jmag@fct.unl.pt, szpektor@google.com

1. Dataset

Each task in the dataset includes a title, a description, a list of ingredients/resources and tools, and a sequence of step-by-step instructions, which may or may not be illustrated. To facilitate the illustration of task steps, we focused on tasks that are mostly illustrated, allowing us to use these images as the ground truth for training and evaluating our methods.

The dataset comprises approximately 1,400 tasks, with an average of 4.9 steps per task, which is a total of 6,860 individual steps. Most tasks include an image for each step, and some feature a complete recipe video that is segmented into multiple clips, with each clip lasting between 10 and 30 seconds per step.

Considering that the number of illustrations can affect the accuracy, we limited the training tasks to those with no more than 10 steps.

2. Model Training

We opted for the CLIP model with a patch size of 32 to serve as the encoder for both image and text data due to its reputation in effectively capturing visual and textual information. In training our own architecture, we conducted experiments with various hyperparameters, including different learning rates, learning rate schedulers, dropout rates, layer freezing, and batch sizes, to identify the most suitable settings for our specific problem.

For the loss function, we employed cross-entropy, comparing the softmax output with hot-encoding of the steps that belong to each task. It is important to note that while this loss function indicates step-task associations, it may not always accurately reflect the model’s overall performance on the task at hand. During the tuning of hyperparameters, we found that freezing layers, weight decay, dropout, and learning rate schedulers had minimal impact on model performance.

The best model, which has about 600,000 parameters,

Training Details	
Optimizer	Adam
Loss Function	Cross-Entropy
Batch Size	500
Learning Rate	0.01
Epochs	10
Model Max Length	400
Number of GPUs	1 A100-40GB

Table 1. Training parameters

was refined using specific training parameters listed in Table 1. Training was completed in under two minutes, using an A100-40GB GPU and spanning ten epochs. Employing the Cross-Entropy loss function, the training process operated with a batch size of 500 and a learning rate set at 0.01, using the Adam optimizer.

Single Modalities. In multimodal generation tasks, the integration of different modalities can notably impact the final output. Through this ablation study, we explore the implications of using singular modalities—text, images, or perturbed inputs—and examine the importance of modality mixing for enhancing generation quality.

Initially, in the scenario where text remains static across inputs, the model struggles with adaptability and generalization due to its reliance on a singular textual context. Conversely, when all inputs are randomized, the absence of consistent patterns across modalities impedes the model’s learning process, resulting in suboptimal performance. However, the configuration where only text is randomized exhibits superior performance, suggesting that the model relies more on image over text. Notably, the marginal difference in performance between random text and the standard training approach underscores the intricate nature of multimodal

tasks.

Our analysis underscores the importance of modality mixing in enhancing multimodal generation tasks. Integrating multiple modalities empowers the model to leverage diverse information sources, leading to more nuanced and accurate outputs. In conclusion, the complexity of multimodal data show that a model can rely more on one modality over the other but a mix of both will always be a better conjunction over a single modality.

Prompt Rewriter Training. The training process was centered on enabling the Large Language Model (LLM) to function as a visual caption generator for original task steps. Leveraging the capabilities of InstructBLIP, we created contextual captions corresponding to each image and its associated step within the dataset. By integrating relevant task context into the generation of ground truth data, we enhanced the LLM’s performance for visual clues. This approach ensured the production of accurate and contextually aligned visual descriptions, solidifying its role as an adept image caption generator.

3. Human annotations

The human annotation process conducted via Amazon Mechanical Turk evaluated multi-scene generated videos. Annotators rated visual quality, entity and background consistency, and text adherence using detailed guidelines. The process included instruction, qualification, and final evaluation phases, comparing our method’s performance against other models and ground truth.

3.1. Annotations Job

Participants for this evaluation were recruited through the crowdsourcing platform Amazon Mechanical Turk. Annotators were compensated at a rate of \$0.5 per task, and each task was designed to take between 2 and 3 minutes to complete.

The payment rate of \$0.5 per task was determined based on pilot tests to estimate the average time required for completion and to ensure fair compensation for participants’ time and effort. At this rate, annotators could earn approximately \$10 per hour if tasks were completed consistently within 3 minutes each, which exceeds the current federal minimum wage in the United States.

All annotators were aware that they were collaborating with researchers for an evaluation on video generation. They received detailed information about their tasks and how their evaluations would contribute to the research.

With focus on the task itself, we maintained annotators’ anonymity. Consequently, we do not have specific demographic or geographic information about the annotators.

3.2. Annotation Process

The annotation process consisted of three main steps:

1. **Instruction Phase:** Annotators received a slideshow with detailed instructions on how to perform the annotations. This phase included several examples to train the annotators and ensure clarity regarding the task requirements.
2. **Qualification Phase:** After the instruction phase, annotators completed a qualification task involving five example annotations. This step was designed to assess their understanding and ability to perform the tasks according to our standards. Only those who passed this qualification phase proceeded to the final annotation phase.
3. **Annotation Phase:** Qualified annotators were then given the full set of annotation tasks, where they evaluated the final results.

3.3. Annotation Tasks

The human annotation pool consisted of annotators who successfully passed the qualification phase.

Figure 1 illustrates the task layout for selecting the best visual coherence maintaining method. Annotators evaluated six models: *CoSeD + Stable Video Diffusion*, *CoSeD + Lumiere*, *TALC + ModelScope*, *TALC + Lumiere*, *Lumiere*, and *Stable Diffusion + Stable Video Diffusion*. They graded the videos on visual quality, entity consistency, background consistency, and adherence to text. The same annotation method was used to compare our best method with other methods as seen in Figure 2.

Figure 3 outlines the annotation guidelines for evaluating the sequences generated by our method compared to other baselines. This figure provides detailed criteria for the annotators to follow, ensuring consistency in their assessments.

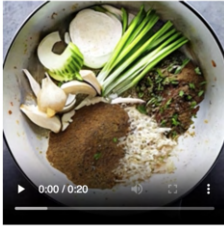
In Figure 4, the specific task of rating sequences generated by our method against ground-truth images is depicted. Annotators were asked to score these sequences on a scale from 1 to 5, providing a quantitative measure of our model’s performance.

To complement this, Figure 5 presents the detailed guidelines used for rating sequences generated by our method compared to ground-truth images. These guidelines helped standardize the evaluation process, ensuring that the ratings were fair and consistent across different annotators.

4. Prompt Optimization

We attempted to enhance generation quality by refining our prompts, incorporating detailed descriptions. These descriptions included:

Answer the following questions based on the multiple-scene descriptions and the candidate generated video.



Descriptions:

Scene 1. 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.
 Scene 2. Add the Small Shrimp, stir well and cook for another 3 minutes.
 Scene 3. 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.
 Scene 4. After about 10-12 minutes, taste for salt and adjust accordingly.

Does the video **scene** exhibit a good **Visual Quality**? (are there any disappearing objects, deformed objects and undesirable artifacts in the video?)

Yes Partially No

Does the video exhibit **Entity Consistency** between **scenes**? (entities are consistent e.g., the shape and features of the objects do not change unless specified)

Yes Partially No

Does the video exhibit **Background Consistency** between **scenes**? (background is consistent when required e.g., the tool does not change without a change described in the scene description)

Yes Partially No

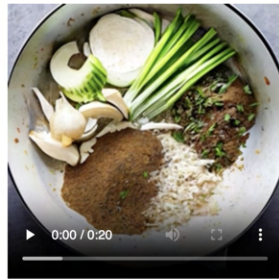
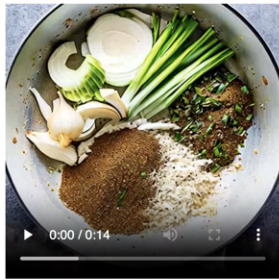
Does the video exhibit **Text Adherence**? (each **scene** is aligned with the textual description)

Yes Partially No

Submit

Figure 1. Human Annotation Layout for Video Generation Methods

Answer the following questions based on the multiple-scene descriptions and the candidate generated video.



Descriptions:

Scene 1. 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.
 Scene 2. Add the Small Shrimp, stir well and cook for another 3 minutes.
 Scene 3. 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.
 Scene 4. After about 10-12 minutes, taste for salt and adjust accordingly.

Which video **scene** exhibits better **Visual Quality**? (are there any disappearing objects, deformed objects, and undesirable artifacts in the video?)

Video 1 Video 2 Both None

Which video exhibits better **Entity Consistency** between **scenes**? (entities are consistent e.g., the shape and features of the objects do not change unless specified)

Video 1 Video 2 Both None

Which video exhibits better **Background Consistency** between **scenes**? (background is consistent when required e.g., the tool does not change without a change described in the scene description)

Video 1 Video 2 Both None

Which video depicts all the **steps** or most of it?

Video 1 Video 2 Both None

Figure 2. Side by Side Annotation Layout for Video Generation Methods

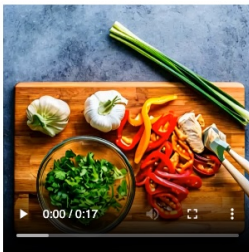
- **Main Subject:** Highlighting the primary focus of the image, whether it's ingredients in a recipe or materials for a project.
 - **Item:** Describing all inanimate objects, ranging from everyday items like utensils or tools to more abstract entities like machinery.
 - **Setting:** Depicting the broader environment or backdrop, spanning from kitchen countertops to workshop benches or outdoor landscapes.
 - **Activity:** Illustrating dynamic actions or steps that animate the imagery, such as stirring ingredients or assembling components.
 - **Arrangement:** Describing the spatial layout, indicating how elements are positioned relative to each other, like 'stacked neatly' or 'arranged in a circular pattern.'
- In the end, though, these prompts failed to produce better outcomes.

Instructions

1. Watch the entire video provided on the left side of the screen.
2. Carefully read the descriptions provided on the right side of the screen.
3. Evaluate the video based on the following criteria:
 - **Visual Quality:** Check if the video scene has good visual quality without any disappearing or deformed objects and no undesirable artifacts.
 - **Entity Consistency:** Ensure that the entities (objects) are consistent between scenes, with no unexpected changes unless specified in the descriptions.
 - **Background Consistency:** Confirm that the background remains consistent between scenes, unless a change is described in the scene description.
 - **Text Adherence:** Verify that each scene in the video aligns with the corresponding textual description.
4. Select the appropriate answer for each question below the video and descriptions.
5. Double-check your answers before submitting the form.

Figure 3. Instructions for Video Generation Evaluation

Rate the video based on the provided descriptions.



Descriptions:

Scene 1. Dice chicken breasts into large pieces. Mince garlic and cilantro, cut bell peppers into strips, and finely chop scallions.

Scene 2. Heat half of the vegetable oil in a large skillet over high heat. Add chicken pieces and sauté for approx. 5 - 7 min. until golden. Transfer to a plate and set aside.

Scene 3. Heat other half of vegetable oil and add scallions and garlic. Sauté until soft and fragrant for approx 1 - 2 min. Then, add bell pepper. Continue to sauté and add corn, chicken stock, and enchilada sauce. Season with salt and pepper.

Scene 4. Bring to a simmer and add rice. Cover and cook on medium-low heat for approx. 15 min. for white rice and 30 min. for brown rice until rice is al dente and most of the liquid absorbed. Preheat oven to 180°C/350°F and turn on the top heat, or set the broiler to high.

Scene 5. Shred chicken and combine with rice. Spread cheese on top, place skillet in the oven, and broil for approx. 5 - 10 min., or until cheese begins to melt. Garnish with cilantro and

Rate the overall quality of the video considering all the mentioned criteria (1 being the lowest and 5 being the highest):

1 2 3 4 5

Submit

Figure 4. Human Annotation Layout for Our Method vs. Ground Truth

5. Selecting the First Image

In the process of generating visual representations based on textual input, the selection of the initial image or video is crucial. This selection not only serves as the first interaction with the user but also influences subsequent representations, directly impacting the overall quality of the generated content. Therefore, establishing a robust strategy for selecting the first image is essential to ensure coherence and effectiveness in the generated output.

The significance of the initial image choice lies in its potential to enhance user engagement. A mismatch between the text and visual representation can disrupt comprehension and decrease the overall user experience. Therefore, the selection strategy should consider factors such as alignment with the text, diversity, and relevance to ensure a seamless

transition from text to visuals.

Single Image Generation. This strategy offers simplicity and directness as its main advantages. By generating a single image, it provides a straightforward solution without added complexity. However, it may suffer from a lack of variety, potentially resulting in limited diversity in the initial representation. Additionally, its reliance on the Stable Diffusion model’s capabilities means that the quality and text adherence of the generated image depends solely on the model’s performance.

Random Selection from Image Batch. The random selection strategy offers increased diversity and reduced bias. By allowing the selection of a random image from a batch,

Instructions

We will present you with a video clip representing a sequence of steps.
Your task is to rate the video on a scale of 1-5 based on the following factors:

- **Representation of Instructions:** How well does the video illustrate the given instructions?
 - *Note:* Any generation artifacts should not impact the rating if the overall video clearly conveys the steps.
- **Coherence:** How coherent is the sequence of scenes in the video?
 - *Example:* If an object is blue in one scene, it should remain blue in subsequent scenes.
 - *Example:* The background should remain consistent across the video.

Figure 5. Instructions for Ground Truth Annotation

it potentially offers a wider range of visual representations. Moreover, it avoids intentional or unintentional bias in selecting the first image. However, it lacks control over the selection process, which may lead to the choice of an image that does not align well with the text. Furthermore, the quality and relevance of the selected image may vary across different runs, introducing potential inconsistency.

Using CLIP for Selection. In this approach, CLIP is used to meticulously select the initial image based on its semantic similarity to the textual description. By leveraging CLIP’s robust understanding of semantics, we ensure that the chosen image corresponds well with the text, thereby enhancing both coherence and relevance. Importantly, the computational overhead associated with CLIP is minimal compared to the resource-intensive task of image generation. Unlike other options, utilizing CLIP for image selection notably increases the likelihood of achieving coherence and relevance in the first generated content.

as the most promising. This strategy emphasizes semantic alignment, ensuring coherence and relevance between the text and visual representation. By leveraging CLIP’s capabilities, we aim to enhance the overall quality and effectiveness of the generated output.

6. Generation Examples

In the following pages, we present several examples of video keyframes and image sequences generated to illustrate specific tasks such as do-it-yourself and recipes using our method and the baselines.



Figure 6. CLIP Selection

Selecting the first image during the generation of visual representations from sequential text input is a critical step. Among the presented strategies for selecting the initial image, the approach of using CLIP for selection stands out

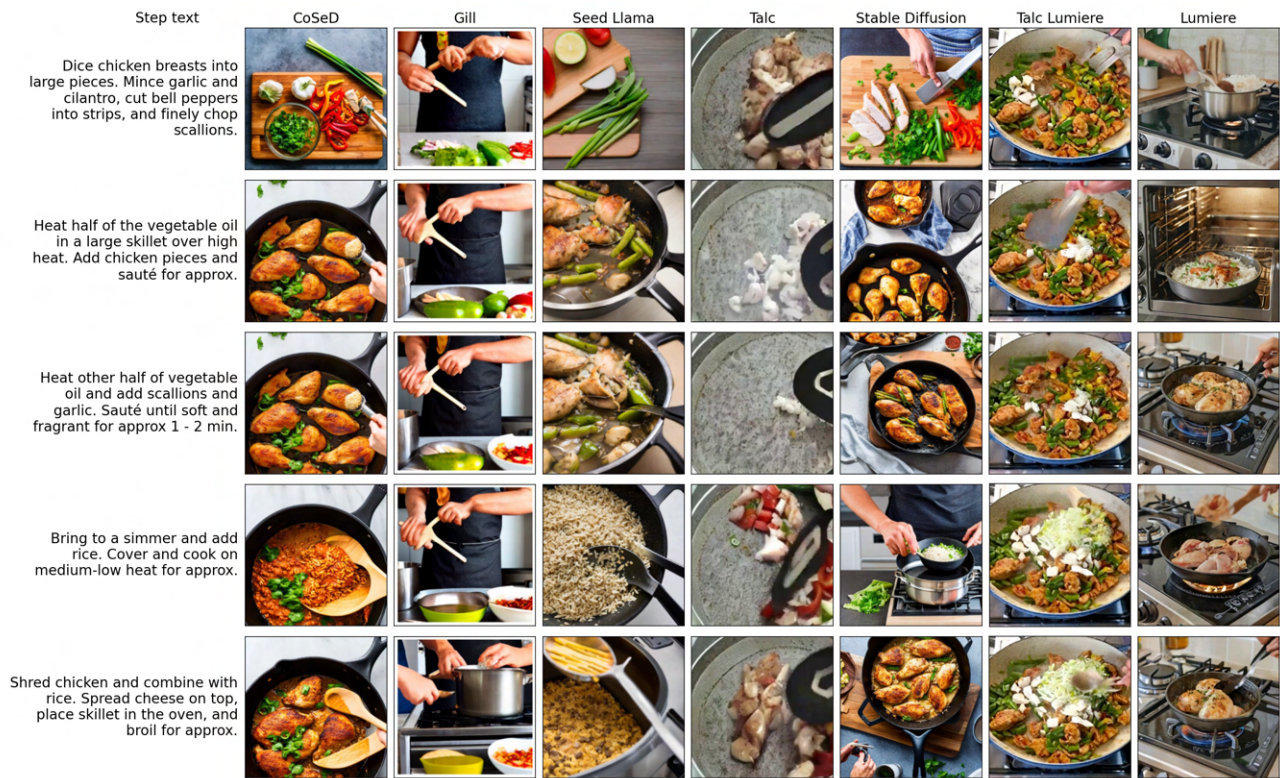


Figure 7. Example of generation with the baselines.

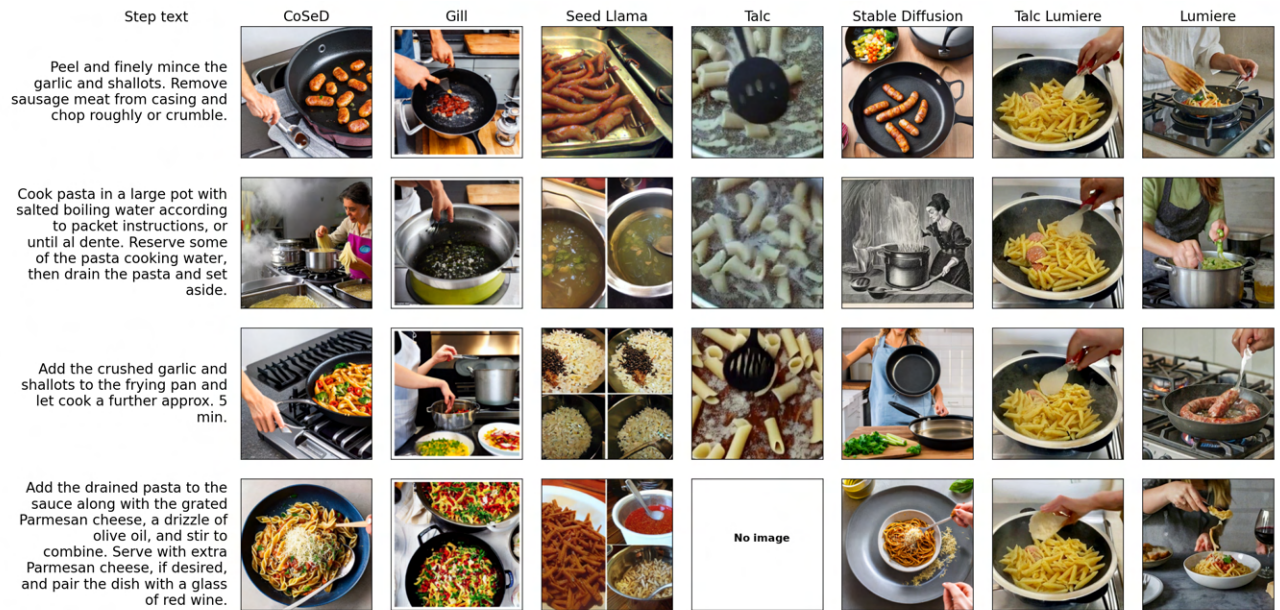


Figure 8. Example of generation with the baselines.

Step text	CoSeD	Gill	Seed Llama	Talc	Stable Diffusion	Talc Lumiere	Lumiere
Melt butter in a small pot over medium heat..							
Add egg and sugar to a bowl and whisk until combined. Stir in honey and add lemon zest.						No image	
Sieve flour and baking powder into the batter, then add melted butter and stir to combine. Cover bowl with plastic wrap and let rest in the fridge for at least 3 hrs.				No image		No image	
Pre-heat oven to 210°C/410°F. Grease baking pan with butter.				No image		No image	
Reduce baking temperature to 170°C/350°F and bake madeleines for approx. 12 - 15 min.				No image		No image	

Figure 9. Example of generation with the baselines.

Step text	CoSeD	Gill	Seed Llama	Talc	Stable Diffusion	Talc Lumiere	Lumiere
Preheat oven to 180°C/350°F and grease springform pan with coconut oil. For the crust, pulse oats in a food processor to a fine flour.							
Use hands to press two-thirds of the dough evenly into springform pan along the base and border. Lightly flour work surface and remaining dough.							
For the filling, peel and quarter apples. Remove core and slice.						No image	
In a large bowl, thoroughly mix apple slices, some cinnamon, nutmeg, lemon juice, agave syrup, arrowroot flour, coconut sugar, speculaas spice, and salt..						No image	
Transfer filling to prepared springform pan and spread evenly..				No image		No image	
Cover filling with either with cookie cut-outs or a round of dough. For the latter, pierce top with a fork to allow steam to escape while baking.				No image		No image	

Figure 10. Example of generation with the baselines.

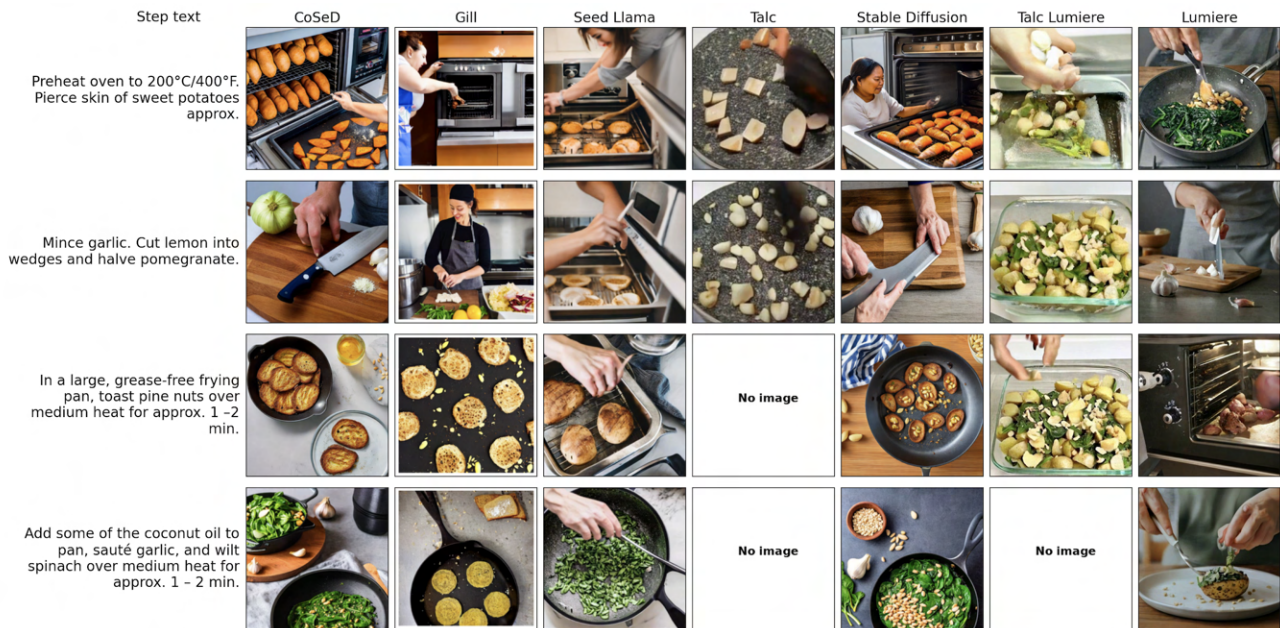


Figure 11. Example of generation with the baselines.



Figure 12. Example of generation with the baselines.



Figure 13. Example of generation with the baselines.

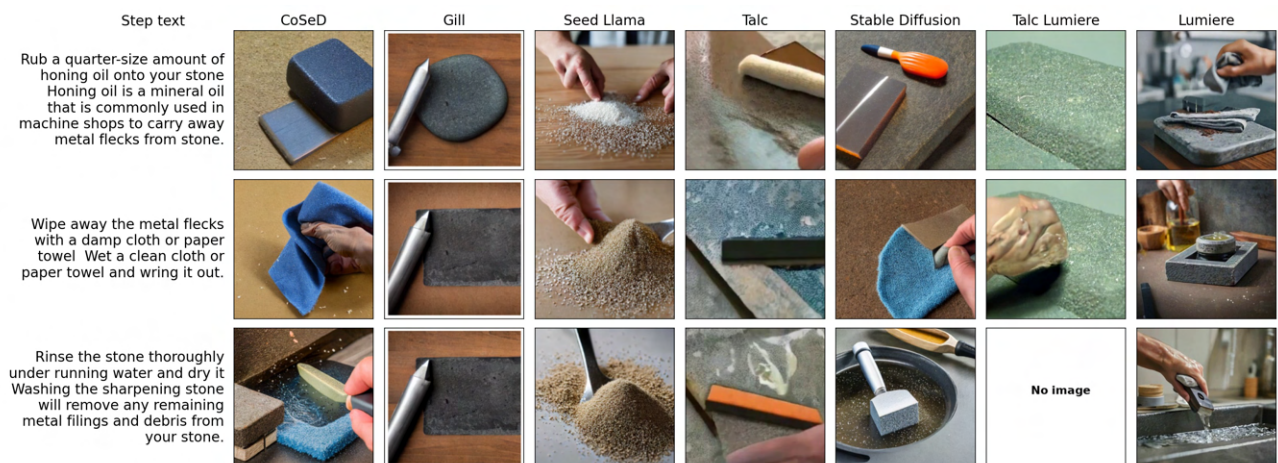


Figure 14. Example of generation with the baselines.



Figure 15. Example of generation with the baselines.

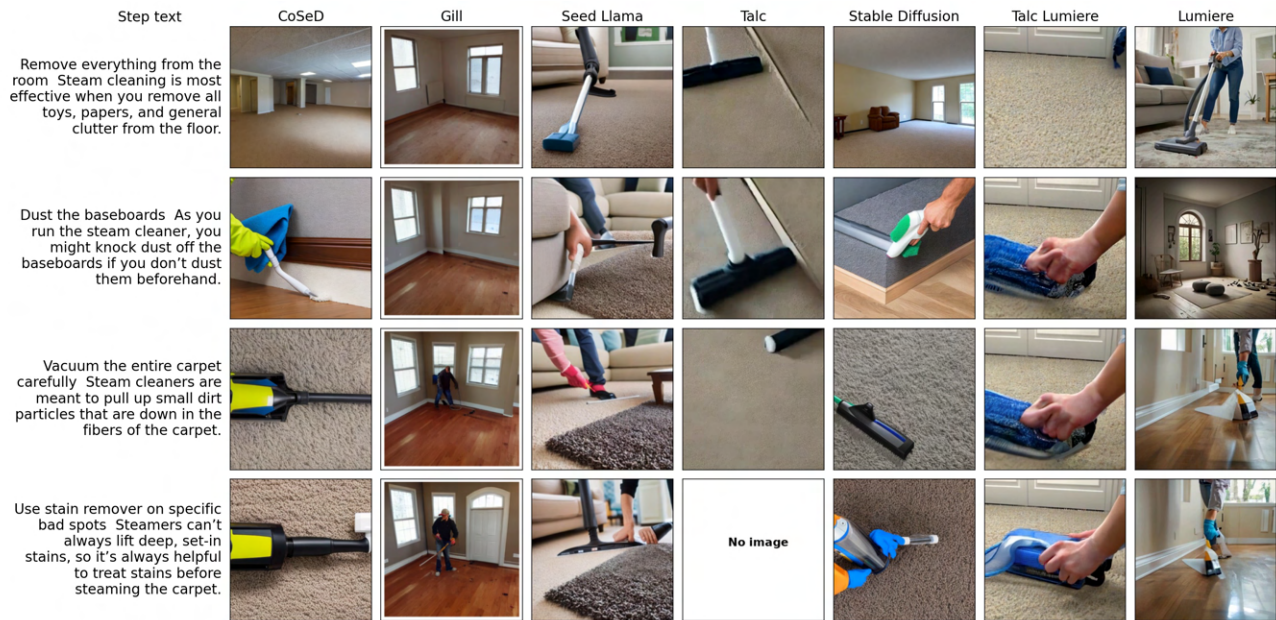


Figure 16. Example of generation with the baselines.

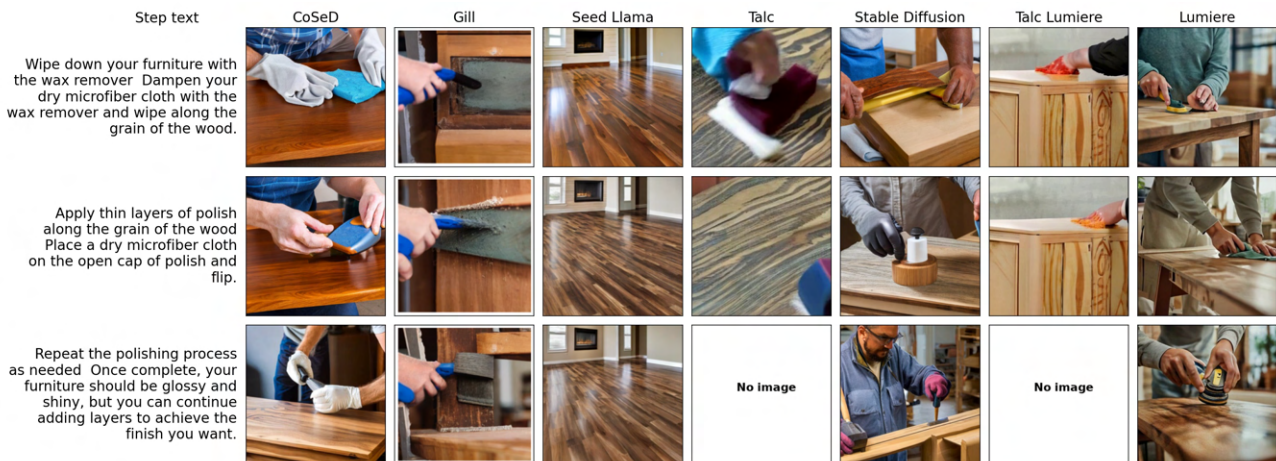


Figure 17. Example of generation with the baselines.