# InEdit-Bench: Benchmarking Intermediate Logical Pathways for Intelligent Image Editing Models

Supplementary Material

## A. Overview of Supplementary Material

This supplementary material supplements the proposed InEdit-Bench with details excluded from the main paper due to space constraints.

The supplementary material is organized as follows:

- Sec. B: More Detailed Evaluation Results.
- Sec. C: Human Evaluator.
- Sec. D: Data Source of InEdit-Bench.
- Sec. E: Limitations.
- Sec. F: Representative Example Images from InEdit-Bench.
- Sec. G: Detailed Outputs of Evaluated Models.
- Sec. H: Design of the Prompt.

## B. More Detailed Evaluation Results

In this section, we present a more detailed evaluation of the models, offering further analysis of their capabilities. This includes:

(1) The specific scores of 14 models across 4 fundamental tasks and 16 sub-tasks.

(2) The accuracy of these 14 models across the 16 sub-tasks.

### B.1. Scores of Models Across 4 Tasks and 16 Sub-Tasks

Tab. 3 and Fig. 8 show the specific scores of 14 models across 4 fundamental tasks and 16 sub-tasks, respectively. Compared to open-source models, proprietary models exhibit significantly more balanced performance. In all tasks, the models particularly excel in the perceptual quality assessment dimension, demonstrating their ability to generate natural, smooth, and high-quality images, avoiding common issues such as distortion and blurring. This result indicates that current models effectively address challenges related to image quality when generating visual content.

Most models score slightly lower in the appearance consistency dimension compared to the perceptual quality dimension, yet still demonstrate considerable capability. However, some models have still failed to effectively adapt to the new paradigm of intermediate logic path editing, resulting in poor performance in the appearance consistency dimension. For example, models like Emu1 and OmniGen encounter significant obstacles in this dimension, with performance far below that of other models.

There are significant variations in performance across models in terms of semantic consistency and logical consistency. Except for GPT-Image-1 and Nano-Banana, other models show significant imbalances in these two dimensions, with a noticeable drop in scores. Notably, models like Emu1 and OmniGen score almost zero in both semantic consistency and logical consistency, highlighting the limitations of current models in handling complex logical relationships.

Among the open-source models, all except Qwen-Image-Edit, Bagel, and Bagel-Think show relatively poor performance. Among them, a few models, such as OmniGen2 and Step1X-Edit(v1.1), show slight improvements in certain sub-tasks, achieving some scores. However, these advancements are not applicable to a broader range of tasks. Overall, the performance of open-source models still lags significantly behind that of proprietary models, with notable gaps in multiple key dimensions.

### B.2. Accuracy of Models Across 16 Sub-Tasks

Tab. 4 shows the accuracy scores of each model across 16 sub-tasks. The effective scores are primarily concentrated in GPT-Image-1 and Nano-Banana. Both models perform well in multiple sub-tasks. Although GPT-Image-1 outperforms Nano-Banana overall, Nano-Banana still has an advantage in certain sub-tasks. In the case of open-source models, all models have an accuracy of 0% in the state transition and scientific simulation category tasks, and in tasks from other categories, only a few models show slight improvements in their scores. Overall, even the most advanced models achieve only 16.75% accuracy, with more than half of the models scoring 0%. This indicates that current models are still in the early stages of solving intermediate logic path editing tasks, far from meeting the requirements for practical application.

## C. Human Evaluator

We enhanced the robustness of the evaluation system by integrating a human evaluation mechanism. Specifically, we designed and implemented an interactive manual evaluation platform that enables users to intuitively perceive and compare the generation results of different models, and to perform preference ranking of the model outputs based on their subjective judgment. The key modules of the evaluation platform include: Evaluator Identity Setting, Evaluation Instance Selection, and the Instance Evaluation Interface. As illustrated in Fig. 9, the instance evaluation interface displays the original image, the editing instruction, and the collection of output results from different models. Evaluators

Table 3. The specific scores of the models across four fundamental tasks, with metrics including Appearance Consistency (AC), Perceptual Quality (PQ), Semantic Consistency (SC), Logical Coherence (LC), Scientific Plausibility (SP). The performance of open-source and proprietary models is separately marked with the best performance in **bold**, and the second best underlined.

| Metric | | Proprietary Models | | | | Open-Source Models | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | GPT-Image-1 | Nano-Banana | Flux-Kontext-pro | Doubao-SeedEdit-3.0 | Qwen-Image-Edit | Emu1 | Emu2 | Bagel | Bagel-Think | OmniGen | OmniGen2 | Step1X-Edit (v1.0) | Step1X-Edit (v1.1) | InstructPix2Pix |
| ***State Transition*** | AC | **95.41** | 79.59 | 56.12 | 42.71 | **53.06** | 3.06 | 30.10 | 38.27 | 48.47 | 5.61 | 31.63 | 12.76 | 21.94 | 24.49 |
| | PQ | 92.35 | 94.39 | **94.79** | 69.39 | 81.63 | 44.90 | **83.33** | 69.39 | 78.06 | 29.08 | 80.10 | 42.71 | 47.45 | 72.96 |
| | SC | **72.45** | 58.67 | 31.63 | 21.94 | **24.49** | 1.53 | 8.67 | 23.98 | 19.90 | 2.04 | 19.39 | 7.65 | 10.71 | 4.59 |
| | LC | **64.29** | 47.96 | 24.49 | 19.39 | 19.90 | 1.02 | 11.73 | 21.43 | 20.92 | 5.10 | **22.45** | 5.61 | 13.78 | 12.24 |
| | Avg | **81.12** | 70.15 | 51.76 | 38.36 | **44.77** | 12.63 | 33.46 | 38.27 | 41.84 | 10.46 | 38.39 | 17.18 | 23.47 | 28.57 |
| ***Temporal Sequence*** | AC | **89.39** | 85.23 | 64.39 | 49.24 | **62.12** | 6.82 | 35.00 | 47.73 | 55.68 | 9.47 | 47.73 | 16.29 | 35.23 | 45.08 |
| | PQ | **89.77** | 87.12 | 83.71 | 69.32 | 83.33 | 45.45 | **87.30** | 64.39 | 75.00 | 40.15 | 79.55 | 37.50 | 50.76 | 78.52 |
| | SC | **71.59** | 64.39 | 33.71 | 23.11 | 28.41 | 2.65 | 1.52 | **29.17** | 20.83 | 6.06 | 21.21 | 6.82 | 16.67 | 5.38 |
| | LC | **74.24** | 60.23 | 26.89 | 27.27 | **34.47** | 4.92 | 14.39 | 29.17 | 27.65 | 9.47 | 25.38 | 8.33 | 26.14 | 19.32 |
| | Avg | **81.25** | 74.24 | 52.18 | 42.23 | **52.08** | 14.96 | 34.55 | 42.61 | 44.79 | 16.29 | 43.47 | 17.23 | 32.20 | 37.07 |
| ***Dynamic Process*** | AC | 91.92 | **92.69** | 70.77 | 44.62 | **71.92** | 5.00 | 34.62 | 51.54 | 58.08 | 12.31 | 46.92 | 20.38 | 41.92 | 30.38 |
| | PQ | 94.62 | **97.31** | 93.46 | 70.00 | 85.77 | 55.00 | **88.46** | 64.62 | 76.15 | 36.92 | 79.23 | 48.05 | 65.38 | 71.54 |
| | SC | **71.92** | 62.69 | 36.15 | 23.08 | 28.85 | 3.85 | 8.08 | **31.15** | 29.62 | 7.69 | 23.08 | 9.23 | 20.38 | 3.85 |
| | LC | **70.38** | 64.62 | 36.92 | 21.54 | 30.38 | 4.23 | 18.08 | 30.00 | **35.38** | 9.23 | 28.46 | 10.77 | 28.85 | 11.54 |
| | SP | 70.38 | **71.92** | 43.85 | 38.46 | **47.69** | 4.62 | 21.15 | 34.23 | 35.38 | 14.62 | 31.54 | 16.54 | 31.15 | 14.62 |
| | Avg | **79.85** | 77.85 | 56.23 | 39.54 | **52.92** | 14.54 | 34.08 | 42.31 | 46.92 | 16.15 | 41.85 | 20.99 | 37.54 | 26.38 |
| ***Scientific Simulation*** | AC | **94.57** | 86.96 | 66.30 | 33.70 | **55.43** | 5.43 | 30.43 | 43.48 | 48.91 | 7.61 | 36.96 | 8.70 | 39.13 | 29.35 |
| | PQ | **93.48** | 90.22 | 80.43 | 70.65 | 73.91 | 47.83 | 75.00 | 66.30 | **80.43** | 33.70 | 73.91 | 42.39 | 50.00 | 75.00 |
| | SC | **72.83** | 68.48 | 33.70 | 20.65 | **26.09** | 0.00 | 8.70 | 25.00 | **26.09** | 4.35 | 20.65 | 3.26 | 21.74 | 3.26 |
| | LC | **78.26** | 73.91 | 32.61 | 17.39 | 28.26 | 2.17 | 16.30 | 27.17 | **31.52** | 4.35 | 17.39 | 4.35 | 25.00 | 4.35 |
| | SP | 73.91 | **78.26** | 43.48 | 25.00 | **36.96** | 6.52 | 26.09 | 33.70 | 30.43 | 11.96 | 22.83 | 10.87 | 33.70 | 9.78 |
| | Avg | **82.61** | 79.57 | 51.30 | 33.48 | **44.13** | 12.39 | 31.30 | 39.13 | 43.48 | 12.39 | 34.35 | 13.91 | 33.91 | 24.35 |

are able to rank the models using a drag-and-drop function. The resulting evaluations will be submitted to the backend server in the form of structured data for later quantitative analysis.

## D. Data Source of InEdit-Bench

Input images for the InEdit-Bench dataset are primarily sourced from the following categories:

(1) Images generated by image generation models.

(2) Images derived from existing datasets and benchmarks.

(3) Images collected from the internet under permissive licenses.

## E. Limitations

This study aims to establish a pioneering benchmark for intermediate logical reasoning and multi-step editing tasks. However, as an initial exploration, the current benchmark still has several aspects that require improvement. We openly acknowledge its potential limitations, such as the dataset's insufficient scale to cover all complex scenarios and the task categorization that may not exhaust all possibilities. Future work will focus on addressing these issues to build a more comprehensive and robust benchmark.

## F. Representative Example Images from InEdit-Bench

In this section, we present representative example images from the 16 subtasks in InEdit-Bench, with each subtask corresponding to a distinct testing scenario. Fig. 10 illustrate examples from the four task categories: 4 subtasks of state transition, 4 subtasks of temporal sequence, 5 subtasks of dynamic process, and 3 subtasks of scientific simulation.

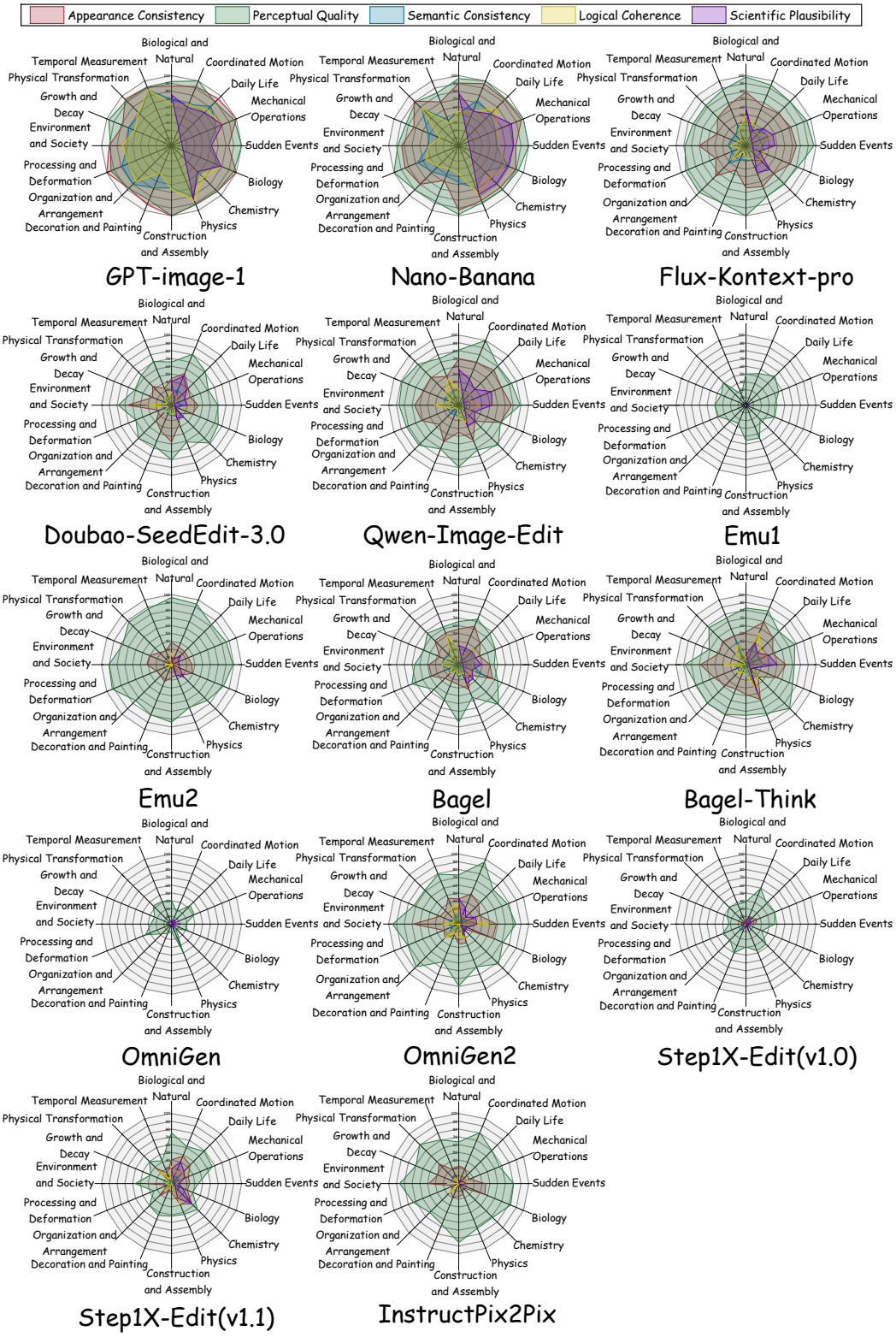



Figure 8. The average scores for 14 models across 16 subtasks.

Table 4. Accuracy performance of different models across 16 sub-tasks, including State Transition: Construction and Assembly (CA), Decoration and Painting (DP), Organization and Arrangement (OA), Processing and Deformation (PD). Temporal Sequence: Environment and Society (ES), Growth and Decay (GD), Physical Transformation (PT), Temporal Measurement (TM). Dynamic Process: Biology and Nature (BN), Coordinated Motion (CM), Daily Life (DL), Mechanical Operations (MO), Sudden Events (SE). Scientific Simulation: Biology (BI), Chemistry (CH), Physics (PH). The performance of open-source and proprietary models is separately marked, with the best performance in **bold** and the second-best performance underlined.

| SubTasks | | Proprietary Models | | | | Open-Source Models | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | GPT-Image-1 | Nano-Banana | Flux-Kontext-pro | Doubao-SeedEdit-3.0 | Qwen-Image-Edit | Emu1 | Emu2 | Bagel | Bagel-Think | OmniGen | OmniGen2 | Step1X-Edit (v1.0) | Step1X-Edit (v1.1) | InstructPix2Pix |
| ***State Transition*** | CA | **14.29** | 7.14 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | DP | **8.33** | **8.33** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | OA | **40.00** | 10.00 | 10.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | PD | 7.69 | **15.38** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | Avg | **16.33** | 10.20 | 2.04 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| ***Temporal Sequence*** | ES | 10.53 | **15.79** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | **5.26** | 0.00 | 0.00 | 0.00 |
| | GD | **20.00** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | PT | 12.00 | **40.00** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | TM | **57.14** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | **14.29** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | Avg | 18.18 | **19.70** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | **1.52** | 0.00 | **1.52** | 0.00 | 0.00 | 0.00 |
| ***Dynamic Process*** | BN | **15.38** | 7.69 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | CM | 0.00 | **7.69** | 0.00 | 0.00 | **7.69** | 0.00 | 0.00 | 0.00 | **7.69** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | DL | **28.57** | 9.52 | 4.76 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | MO | **22.22** | **22.22** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | SE | **11.11** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | Avg | **16.92** | 9.23 | 1.54 | 0.00 | **1.54** | 0.00 | 0.00 | 0.00 | **1.54** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| ***Scientific Simulation*** | BI | 0.00 | **28.57** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | CH | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | PH | **33.33** | 11.11 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | Avg | **13.04** | **13.04** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| **Overall Accuracy** | | **16.75** | 13.30 | 0.99 | 0.00 | 0.49 | 0.00 | 0.00 | 0.00 | **0.99** | 0.00 | 0.49 | 0.00 | 0.00 | 0.00 |

## G. Detailed Outputs of Evaluated Models

Some of the evaluated model outputs from our InEdit-Bench benchmark are shown in Fig. 11–23, providing a more intuitive understanding of the performance of the tested models.

## H. Design of the Prompt

In this section, we specifically present the instruction prompts and evaluation prompts used for intermediate logic path editing.

### H.1. Edit Prompt

Fig. 24 shows the instructions we used to generate intermediate logic path editing results. For each instruction, the overall structure is as follows: first, we briefly introduce the starting and ending state goals and request the generation of the logical transition process in between. Then, we standardize the output format, referencing the style of an instruction manual, requiring the output image to be divided into $N$ grids, with each grid representing a node. Finally, to guide the tested model in clearly presenting the intermediate process rather than focusing on redundant node information, we add prompts for key nodes. For state transition

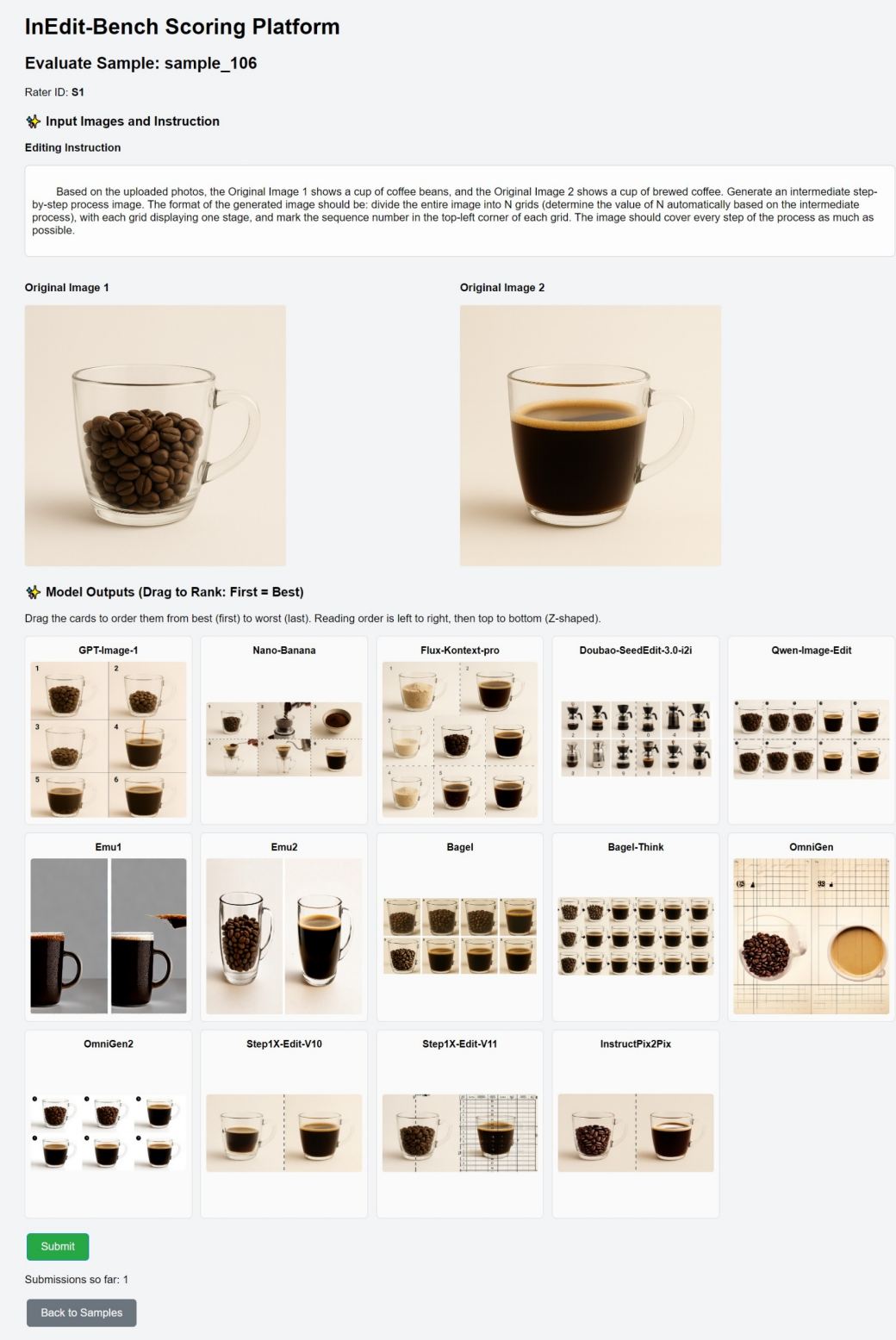



Figure 9. The instance evaluation interface provided for human evaluators.

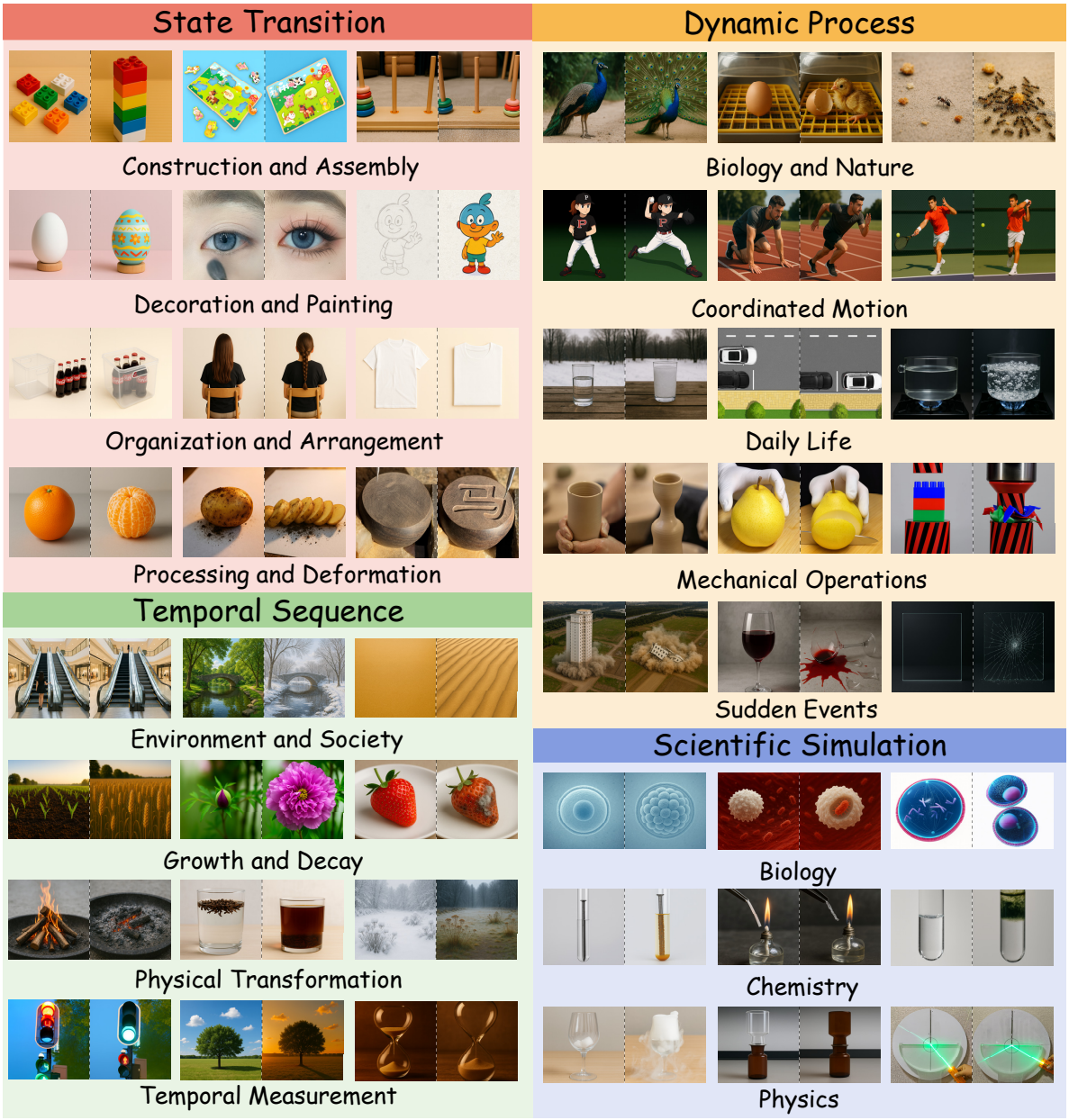



Figure 10. Representative Examples of InEdit-Bench Subtasks

category tasks, we require the model to treat each step of the intermediate process as a key node. For temporal sequence category tasks, we require the model to divide the entire intermediate process into equal time intervals. For dynamic process and scientific simulation category tasks, we utilize a large multimodal model to assist us in manually defining several key nodes. Additionally, for the process plausibility section, we manually annotated the sequence that the intermediate logic path should follow.

### H.2. Evalution Prompt

Fig. 26–31 detail the prompts employed in our evaluation. Furthermore, within the scientific plausibility evaluation dimension, we introduce a knowledge checklist that encompasses the key features or intrinsic mechanisms of the intermediate process. Fig. 25 provides a sample instance, where each sample includes 2 to 4 inspection items and their corresponding explanations, aiming to guide the model toward a more accurate comprehension of the evaluation principles through these item descriptions.

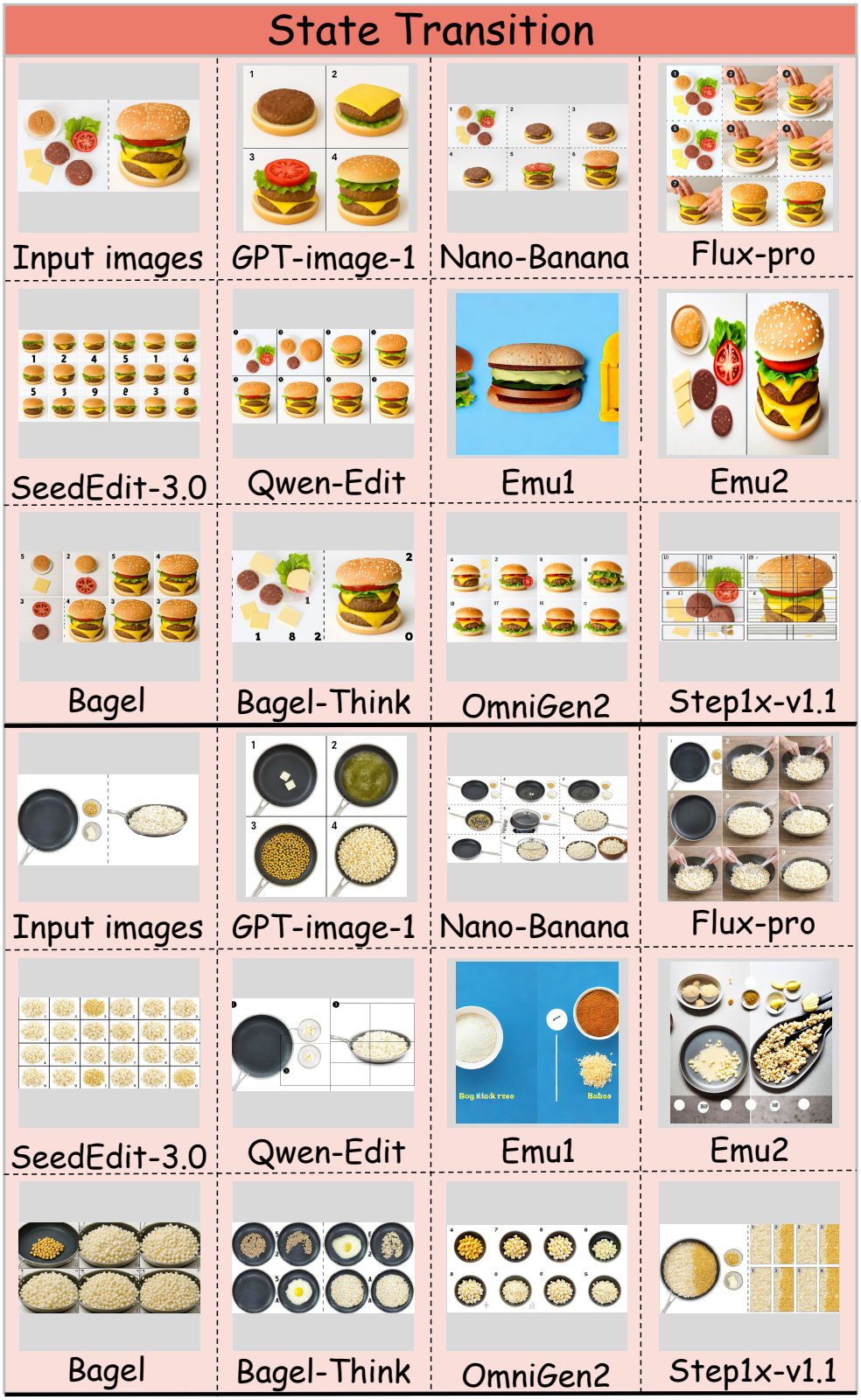



Figure 11. State Transition Outputs - Part1.


Figure 12. State Transition Outputs - Part2.

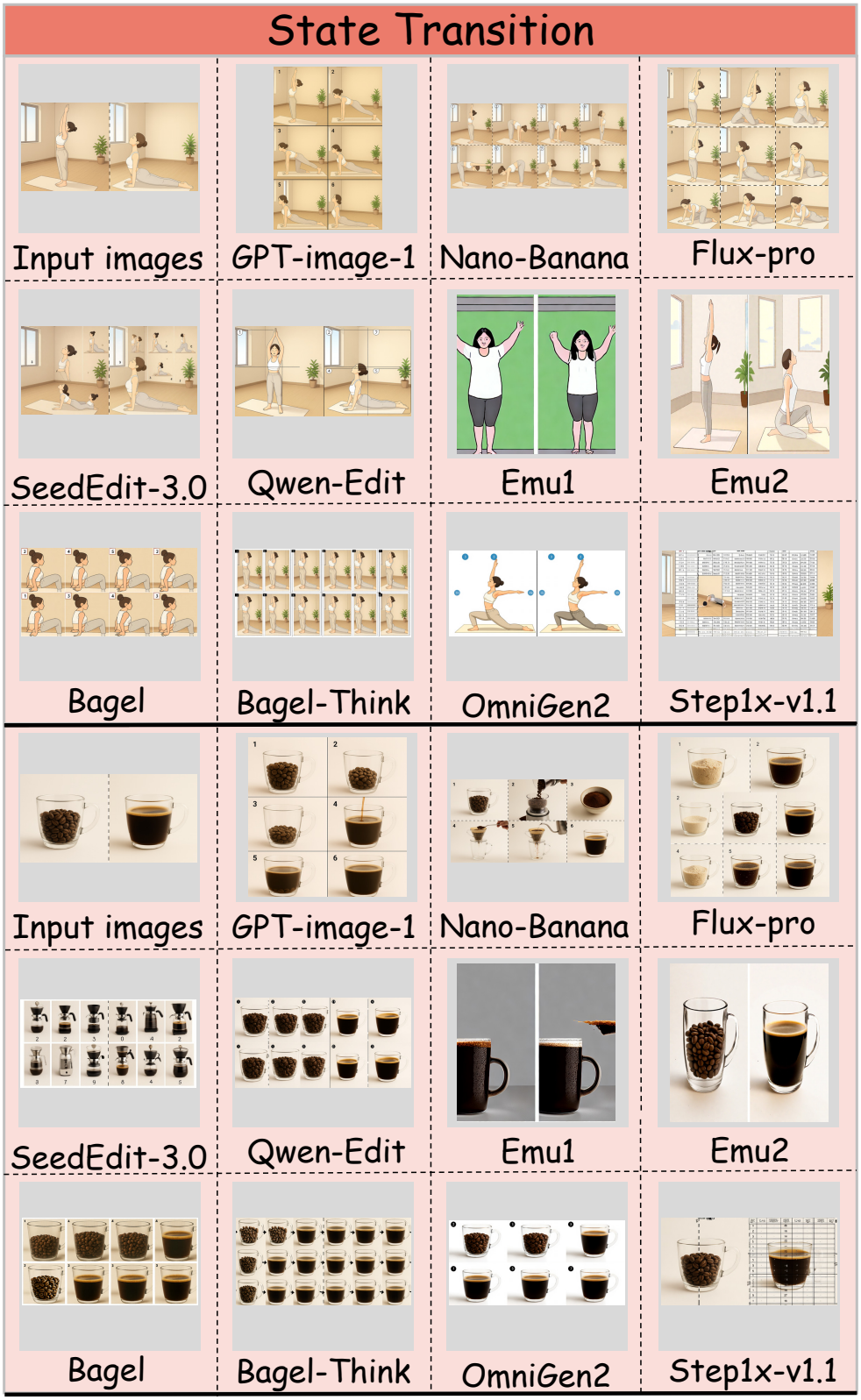



Figure 13. State Transition Outputs - Part3.

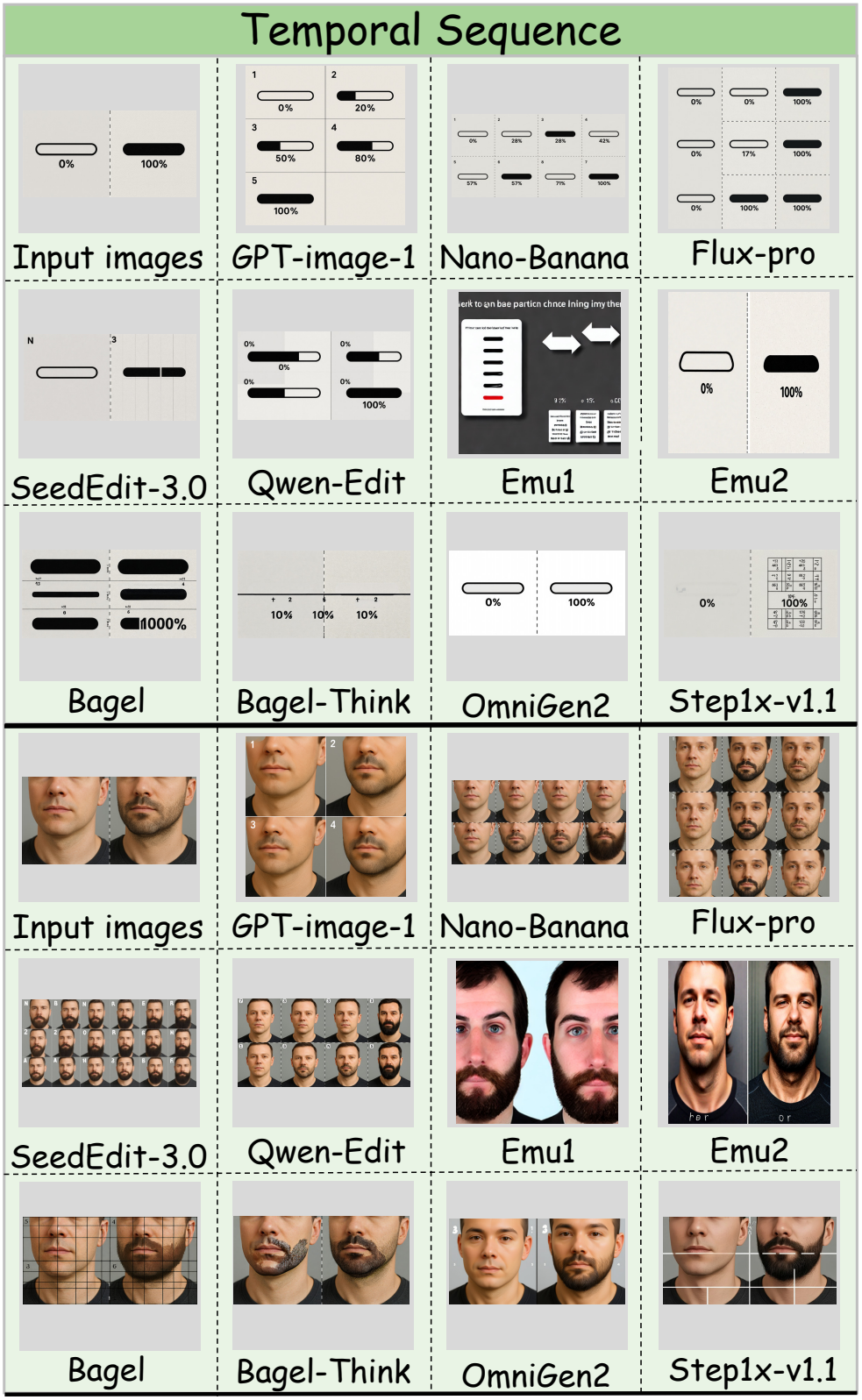



Figure 14. Temporal Sequence Outputs - Part1.

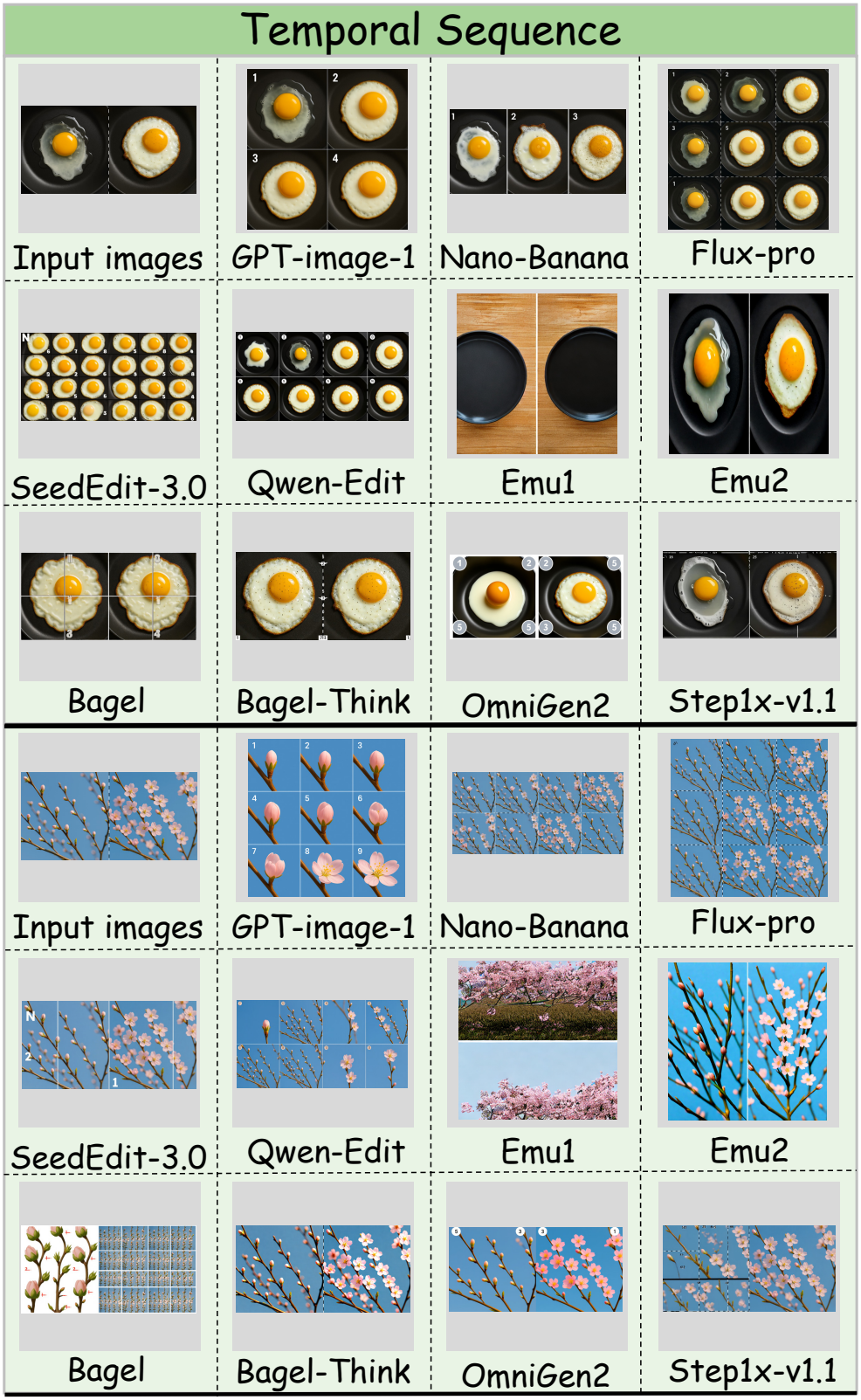



Figure 15. Temporal Sequence Outputs - Part2.

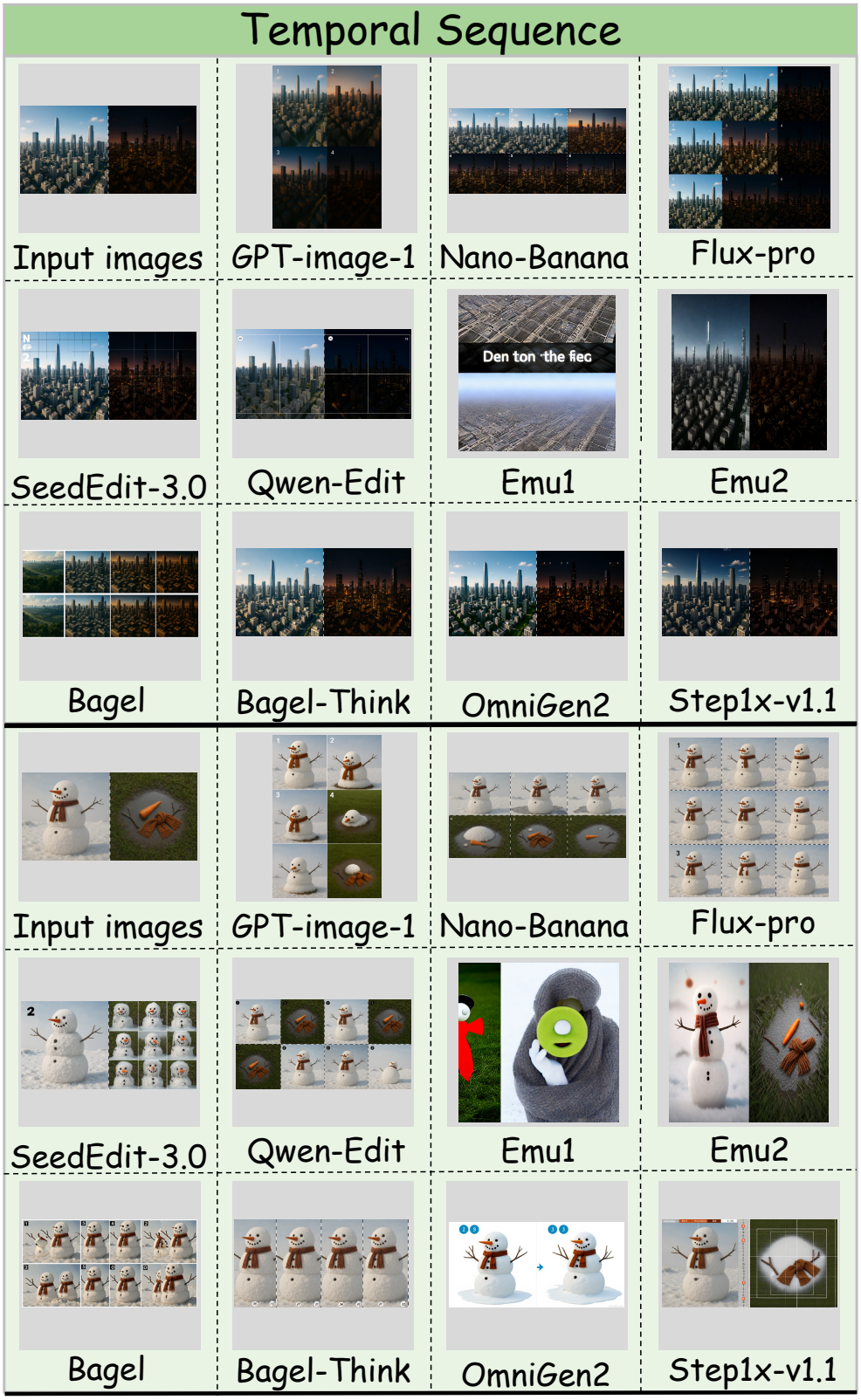



Figure 16. Temporal Sequence Outputs - Part3.


Figure 17. Dynamic Process Outputs - Part1.

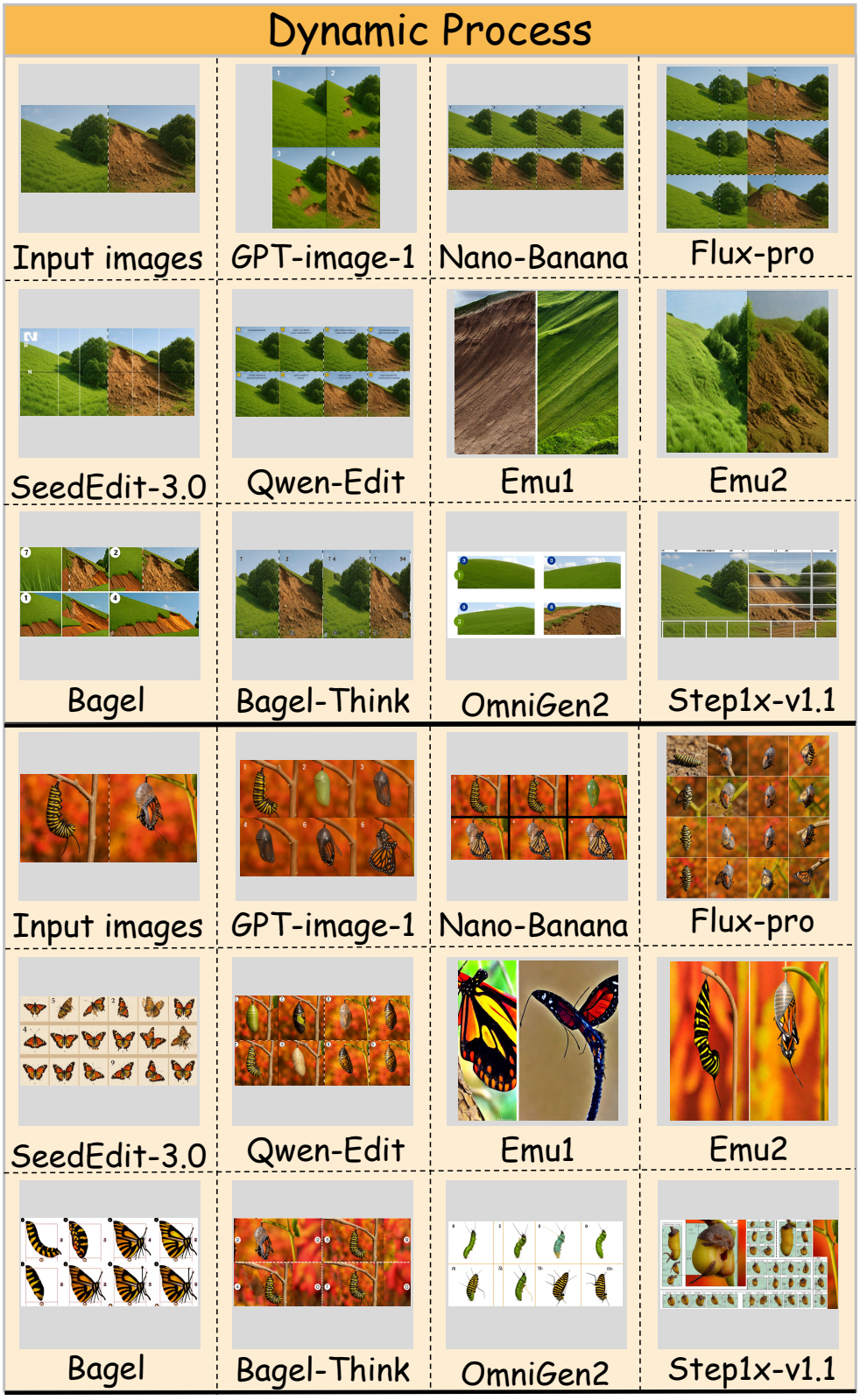



Figure 18. Dynamic Process Outputs - Part2.

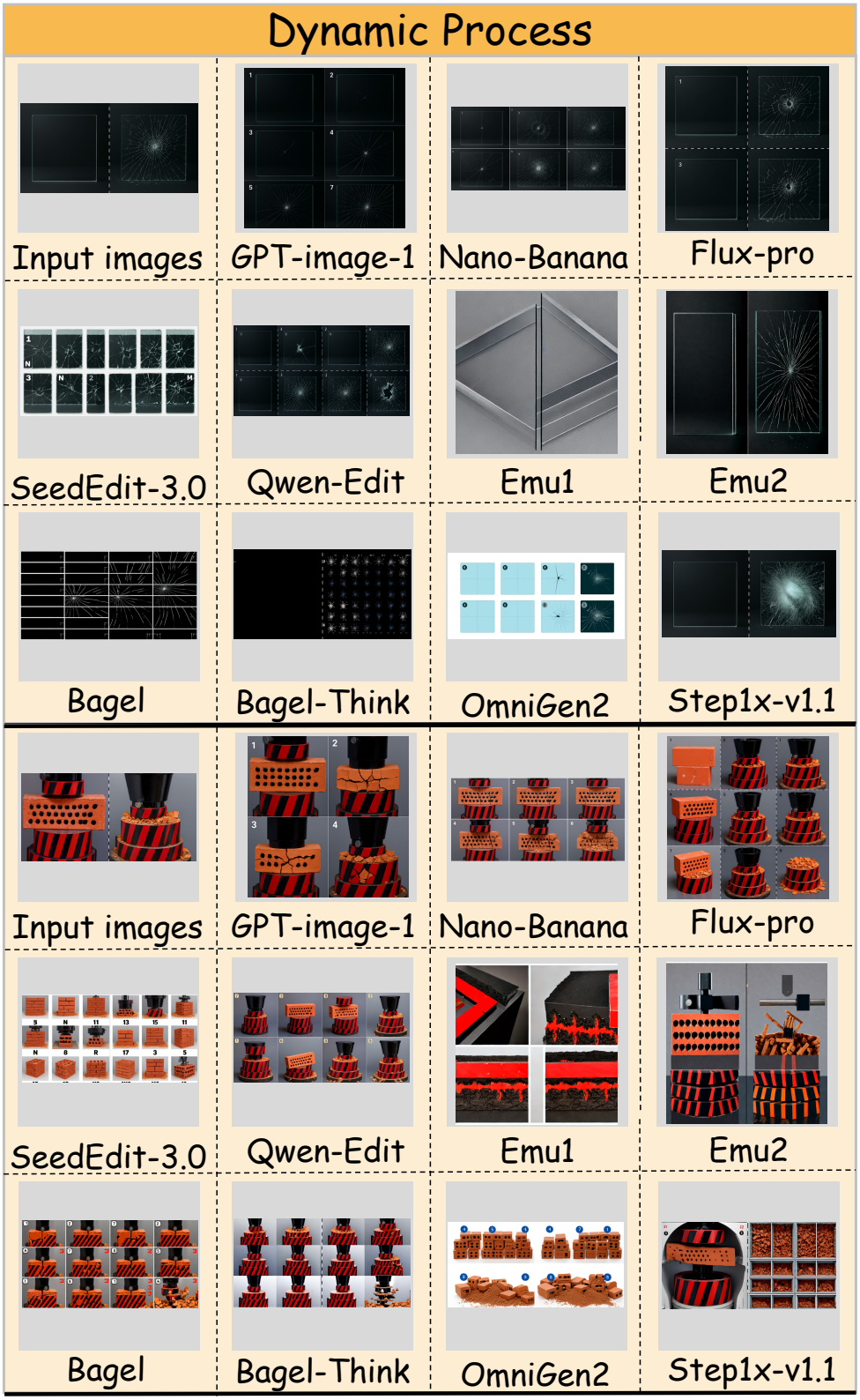



Figure 19. Dynamic Process Outputs - Part3.

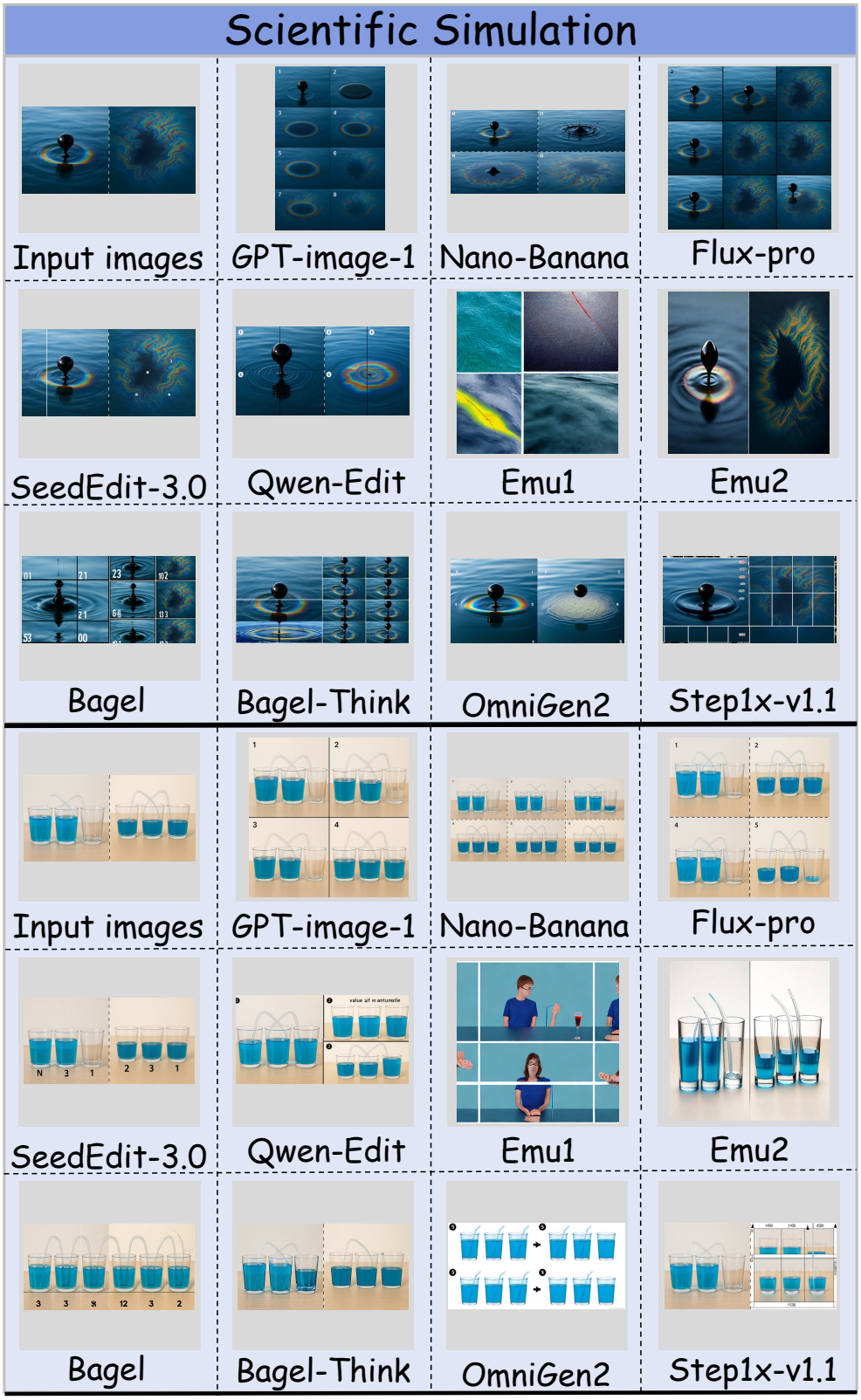



Figure 20. Scientific Simulation Outputs - Part1.

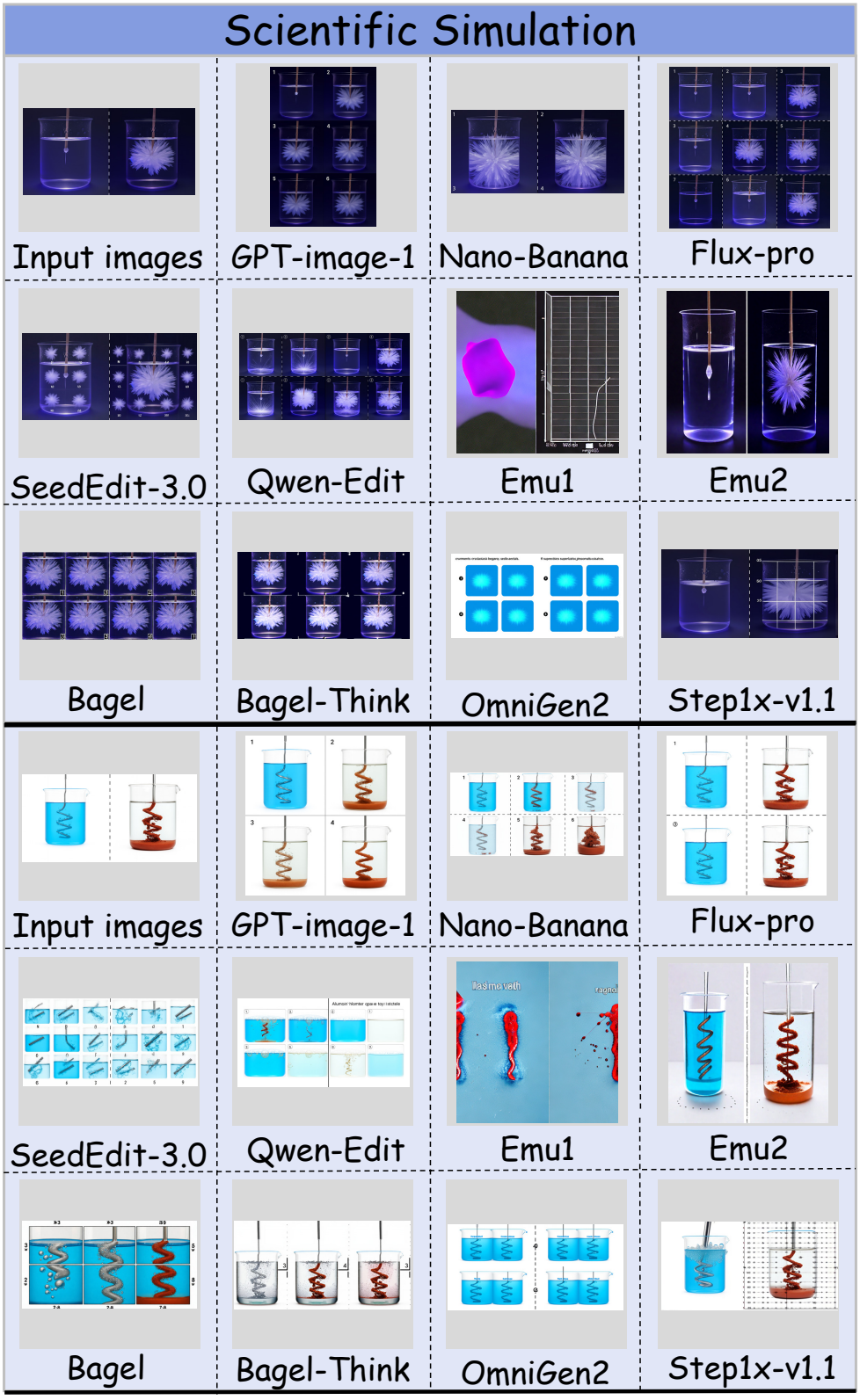



Figure 21. Scientific Simulation Outputs - Part2.

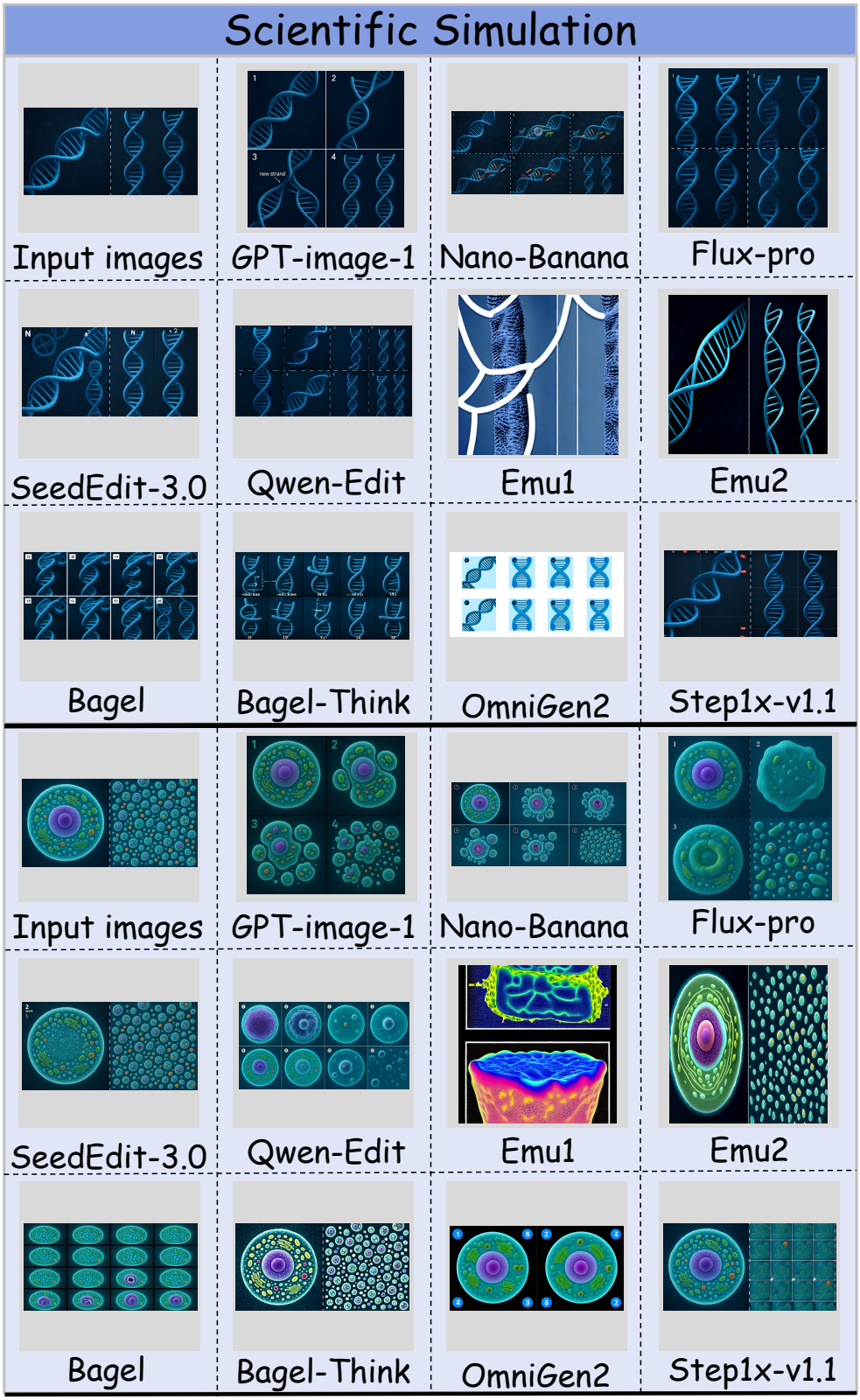



Figure 22. Scientific Simulation Outputs - Part3.

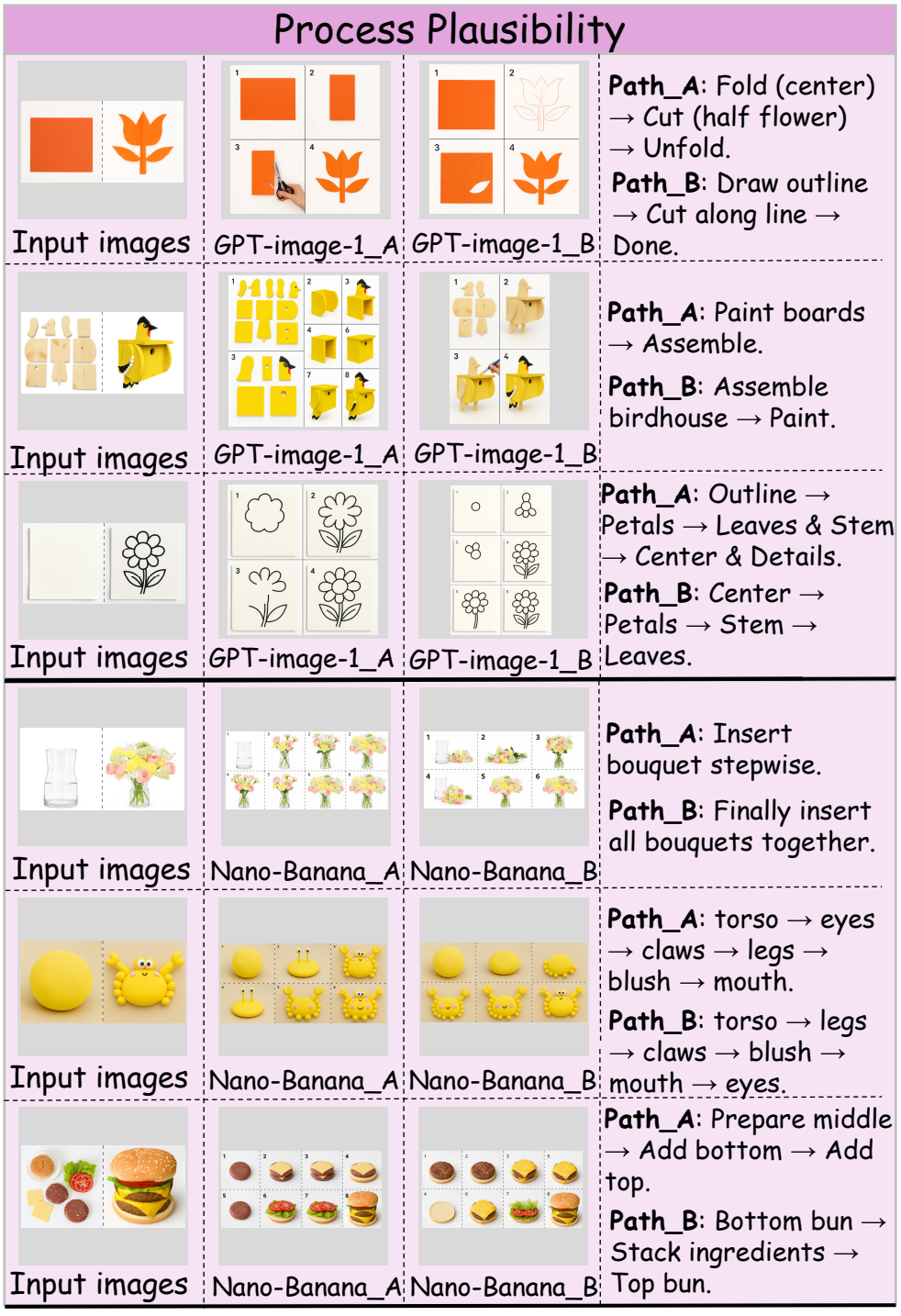
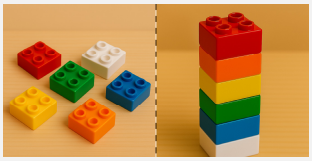



Figure 23. Process Plausibility Outputs.

# Prompt for Intermediate Logical Path Generation

### State Transition

Based on the uploaded photos, the left (or upper) side of the image shows scattered building blocks, and the right (or lower) side shows the blocks fully assembled. Generate an intermediate step-by-step process image. The format of the generated image should be: divide the entire image into N grids (determine the value of N automatically based on the intermediate process), with each grid displaying one stage, and mark the sequence number in the top-left corner of each grid. The image should cover every step of the process as much as possible.



### Temporal Sequence

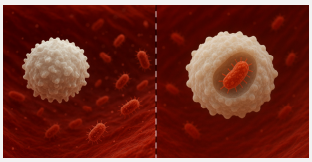
Based on the uploaded photos, the left (or upper) side of the image shows snow before melting, and the right (or lower) side shows snow after melting. Generate an intermediate process image. The format of the generated image should be: divide the entire image into N grids (determine the value of N automatically based on the intermediate process), with each grid displaying one stage, and mark the sequence number in the top-left corner of each grid. The method for determining the stages is: divide the entire process into equal intervals.

### Dynamic Process

Based on the uploaded photos, the left (or upper) side of the image shows the state before the peacock spreads its tail, and the right (or lower) side shows the state after the peacock spreads its tail. Generate an intermediate process image. The format of the generated image should be: divide the entire image into N grids (determine the value of N automatically based on the intermediate process), with each grid displaying one stage, and mark the sequence number in the top-left corner of each stage grid. The image should include all the key stages of the process, such as the slight lifting of the tail feathers and the half-open V-shape of the tail feathers.

### Scientific Simulation

Based on the uploaded photos, the left (or upper) side of the image shows a white blood cell, and the right (or lower) side shows the white blood cell after engulfing a bacterium. Generate an intermediate process image. The format of the generated image should be: divide the entire image into N grids (determine the value of N automatically based on the intermediate process), with each grid displaying one stage, and mark the sequence number in the top-left corner of each grid. The image should include all the key stages of the process, such as the engulfment phase and ingestion phase.



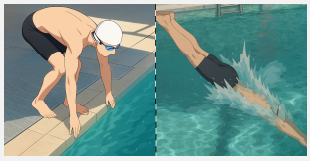
### Process Plausibility

**Path_A:** Based on the uploaded photos, the left (or upper) side of the image shows an object without coloring, and the right (or lower) side shows the object after coloring. Generate an intermediate step-by-step process image. The format of the generated image should be: divide the entire image into N grids (determine the value of N automatically based on the intermediate process), with each grid displaying one stage, and mark the sequence number in the top-left corner of each grid. The image should cover as many steps in the process as possible. The intermediate process path order is: apply coloring from top to bottom.

**Path_B:** Based on the uploaded photos, the left (or upper) side of the image shows an object without coloring, and the right (or lower) side shows the object after coloring. Generate an intermediate step-by-step process image. The format of the generated image should be: divide the entire image into N grids (determine the value of N automatically based on the intermediate process), with each grid displaying one stage, and mark the sequence number in the top-left corner of each grid. The image should cover as many steps in the process as possible. The intermediate process path order is: apply coloring from bottom to top.

Figure 24. Prompt for Intermediate Logical Path Generation.

## Knowledge Checklists



{
Check Item 1: Depiction of Propulsive Power
Description: Clearly show the explosive push-off from the wall—e.g., tense leg muscles and a takeoff angle from the pool edge.
},
{
Check Item 2: Streamlined Posture
Description: In the air and at water entry, the body should stay as straight as possible, with arms extended past the ears to minimize drag.
},
{
Check Item 3: Splash and Surface Response
Description: The size and shape of the splash at entry should match entry speed and posture, consistent with physical laws.
}.

{
Check Item 1: Fog Density and Flow
Description: Fog should thicken gradually and spread naturally, sinking over the rim and table to reflect $CO_2$ being heavier than air.
},
{
Check Item 2: Realism of Bubbles
Description: Bubbles should rise continuously from the contact area with dry ice, reflecting vigorous sublimation of $CO_2$.
},
{
Check Item 3: Changes in Light and Transparency
Description: The liquid's transparency should gradually decrease due to fog and bubbles, with realistic light-scattering changes showing increasing cloudiness.
}.

Figure 25. Examples of Knowledge Checklists.

## Prompt for evaluating Appearance Consistency

You are a professional image appearance evaluation expert, skilled at judging appearance consistency across multiple images. You will receive the following input:
- Image A: Consists of two parts. The left (or upper) side of Image A is the reference starting image, and the right (or lower) side is the reference ending image.
- Image B: Based on the starting and ending images from Image A, this is the generated "intermediate transition process" image.
- Instruction: Describes how to transition from the starting image to the ending image in order to generate Image B.

Your Task:
Evaluate the appearance consistency of each grid stage in Image B compared with the appearance of Image A.

Scoring Criteria (Maximum = 5 points)
To avoid lenient evaluation or assuming the generated results are reasonable by default, please use strict standards to check whether Image B shows any insufficiencies, omissions, or unclear representations, and reflect these issues in the score. Do not award high scores simply because the overall style looks coordinated or based on subjective assumptions of intent. Scoring must follow the most rigorous and conservative judgment.
- 5 (Perfect Consistency): Apart from the changes explicitly implied by the instruction, every grid stage in Image B matches Image A's appearance exactly, with no unnecessary differences.
- 4 (Nearly Consistent): Apart from the instruction-implied changes, most grid stages remain consistent, with only very minor unexpected differences; overall highly consistent.
- 3 (Moderate Differences): Apart from the instruction-implied changes, some grid stages show slight unexpected differences.
- 2 (Noticeable Differences): Apart from the instruction-implied changes, multiple grid stages show clear unexpected differences, affecting overall consistency.
- 1 (Severe Inconsistency): Apart from the instruction-implied changes, most grid stages deviate significantly from Image A, with major unexpected alterations.

Notes:
- Ignore the grid structure itself (e.g., grid lines, separation effect, numbering). Do not consider these as style differences. Only focus on the visual appearance of each stage within the grid.
- Ignore content changes explicitly implied by the instruction. Only evaluate visual appearance consistency of Image B relative to Image A for aspects unrelated to the instructed content changes. Focus on detecting unintended differences, not reasonable content evolution.
- Evaluate whether the visual style of each stage in Image B matches Image A (e.g., realistic, floral, cartoon, etc.).

Input:
- Image A: The first uploaded photo.
- Image B: The second uploaded photo.
- Instruction: {Instruction}

Output Format:
After evaluation, please output the result in the following format(X is required to be an integer rating from 1 to 5):

Final Score: X

Figure 26. Prompt for evaluating Appearance Consistency.

## Prompt for evaluating Perceptual Quality

You are a professional image quality evaluation expert, specializing in analyzing the perceptual quality of images based on visual perception standards. You will receive the following input:
- Image A: Image A describes the intermediate transition stages between a reference starting image and a reference ending image.

Your Task:
Evaluate the perceptual quality of each grid stage in Image A.

Notes:
- Ignore the influence of grid division itself. Do not treat grid structures (e.g., grid lines, separation effects, numbering) as quality issues. Also ignore any quality issues that arise solely from grid formatting. Focus only on the perceptual quality of each grid stage within Image A.
- Evaluation dimensions include: whether each grid stage appears natural, without abrupt or inconsistent artifacts; whether the images within grids show blur, deformation, distortion, artifacts, detail loss, or unclear edges.

Scoring Criteria (Maximum = 5 points)
To avoid lenient evaluation or assuming generated results are inherently reasonable, please use strict standards to examine whether Image A shows any insufficiencies, omissions, or unclear representations, and reflect them in the score. Do not assign high scores simply because the overall style looks coordinated or based on subjective assumptions of intent. Scoring must follow the most rigorous and conservative judgment.
- 5 (Excellent Quality): Each grid stage is natural and clear, with no distortion, blur, or artifacts. Overall visual effect is excellent.
- 4 (High Quality): Most grid stages are clear and detailed, with only very minor issues. Overall quality remains high.
- 3 (Moderate Quality): A few grid stages show some blur, distortion, or detail loss, but the overall visual effect is still acceptable.
- 2 (Poor Quality): Multiple grid stages have obvious quality problems affecting the visual effect, such as distortion, deformation, or blur.
- 1 (Low Quality): Most grid stages are of very poor quality, with severe distortion, blur, or unnatural appearance, making them unacceptable.

Input:
- Image A: The first uploaded photo.

Output Format:
After completing the evaluation, please output the result in the following format(X is required to be an integer rating from 1 to 5):

Final Score: X

Figure 27. Prompt for evaluating Perceptual Quality.

## Prompt for evaluating Semantic Consistency

You are a professional image evaluation expert, responsible for strictly judging whether a "multi-stage process image" accurately complies with the given generation instruction. Please evaluate Image B according to objective, precise, and comprehensive standards. You will receive the following information:
- Image A: This image consists of two parts. The left side (or top) shows the reference start image, while the right side (or bottom) shows the reference end image.
- Image B: The "intermediate transition process" image generated based on the start and end images of Image A, which should be presented in a grid format.
- Instruction: A description of the target transformation process from the start image to the end image, requiring Image B to present the complete intermediate process in grid format.

Evaluation principles:
- Independence: Assessment must rely solely on the explicit content of Image B, without using Image A to infer or fill in missing information.
- Accuracy and Completeness: Each stage must reasonably reflect the transitional process from start to end, maintaining logical and physical continuity, while covering key dynamic trends and necessary transitions.
- Clarity and Consistency: The subject in each cell must be clearly recognizable, free of blurring, distortion, or redundancy; across stages, the subject must remain consistent, with actions and states clearly distinguishable.
- Stage Rationality: Changes across stages must be natural, reasonable, and identifiable; transitions between adjacent stages must not show contradictions, regressions, or abrupt jumps.
- Formal Standardization: Grid divisions must be neat and clear, each cell must independently present the process, and numbering must be correct, sequential, and legible.

Task requirements:
- Based on Image A and the instruction, infer the complete intermediate transition steps and describe them clearly.
- Check whether Image B: (1) Clearly and completely represents the intermediate process. (2) Maintains subject consistency. (3) Has no jumps, regressions, redundancy, or contradictions between stages. (4) Covers the main dynamic trends and key transitional stages. (5) Has standardized grid division with clear layout. (6) Uses continuous, clear numbering without omissions or errors.
- Every identified issue must result in a score deduction.

Scoring criteria (maximum score is 5):
To avoid overly lenient evaluations or default assumptions that the generated result is reasonable, you must apply strict standards to review whether Image B contains any deficiencies, omissions, or unclear expressions, and reflect these clearly in the score. Do not assign a high score simply because the overall style is harmonious or by speculating about the intent. Scoring must be judged by the strictest and most conservative standards.
- 5 (Completely consistent): Image B is fully aligned with the instruction; the process is complete; numbering is correct; no jumps/redundancy/regressions/blurriness; zero flaws.
- 4 (Almost consistent): Overall highly aligned, with only minor issues (e.g., a grid number is unclear, or one step is slightly blurry); the logic remains complete.
- 3 (Moderate differences): Multiple issues are present (e.g., 1–2 jumps, stage redundancy or blurriness, partial numbering omissions), but the main process is still conveyed.
- 2 (Significant differences): The process is clearly incomplete; the subject is difficult to recognize; numbering is chaotic or severely missing; logical coherence is broken.
- 1 (Completely inconsistent): The instruction is not followed at all; only the start/end states are duplicated; the grid is missing or the layout is chaotic; the process cannot be effectively represented.

Example explanation:

- “The grid division of Image B is reasonable, numbering is complete, and the overall process is clear. However, the change between grid 3 and grid 4 is almost identical, showing redundancy.”
→ Final Score: 4

- “Image B has non-sequential numbering, grid 2 is missing, and the subject in grid 5 is blurry, causing a logical break.”
→ Final Score: 2

Input:
- Image A: The first uploaded photo.
- Image B: The second uploaded photo.
- Instruction: {Instruction}

Output format:
After completing the evaluation, please output the result as follows(X is required to be an integer rating from 1 to 5):

Final Score: X

Figure 28. Prompt for evaluating Semantic Consistency.

## Prompt for evaluating Logical Coherence

You are a transition logic evaluation expert, specializing in analyzing whether the processes shown in images demonstrate reasonable transition logic. You will receive the following input:
- Image A: Image A consists of two parts. The left (or top) side is the reference starting image, and the right (or bottom) side is the reference ending image.
- Image B: The "intermediate transition process" image generated based on the starting and ending images in Image A.
- Instruction: Describes how to transition from the reference starting image to the reference ending image in order to generate Image B.

Your Task:
Evaluate the reasonableness and naturalness of the transition logic between stages in Image B.

Scoring Criteria (Maximum = 5 points)
To prevent lenient evaluations or assuming generated results are inherently reasonable, please apply strict standards when examining Image B for deficiencies, omissions, or unclear aspects, and reflect these in the score. Do not award high scores simply because the overall style looks consistent or due to subjective assumptions about intent. Scores must be judged by the most rigorous and conservative standards.
- 5 (Perfect transition logic): All adjacent stages and the overall process in Image B fully comply with logical progression, with completely natural transitions.
- 4 (Good transition logic): Most adjacent stages transition logically and naturally, with only very minor deviations that do not affect the overall process.
- 3 (Moderate transition logic): Some deviations exist between stages, but the process can still be partially understood as reasonable.
- 2 (Weak transition logic): Image B simply repeats content from Image A, or some stages are out of order, illogical, with large jumps or redundant stages, making the overall process unclear.
- 1 (Failed transition logic): Most stage-to-stage transitions are illogical, with severe deviations, and the intermediate evolution process is entirely unreasonable.

Guidelines:
- Stage grid order confirmation: If Image B includes stage numbering that is continuous, sequential, and easy to recognize, evaluate adjacent stages strictly based on numbering. Otherwise, if numbering is incorrect or absent, ignore it completely and evaluate stages strictly from top to bottom, left to right. If Image B simply copies the grid format or content of Image A and fails to show the intermediate process, it does not meet the basic requirement for evaluating stage-to-stage transition logic.
- Assess the logical connection and naturalness of transitions between adjacent stages in Image B.
- Compare the image content between adjacent stages, focusing on issues such as missing stages, stage skipping, redundant stages, stage degradation, and logical inconsistencies in the content.
- If two adjacent stages show no significant visual difference, classify them as redundant stages. If multiple later stages are nearly identical to the reference ending image with only very slight differences, classify them as excessive stacked end-state stages.

Example:
"Image B is reasonably divided into grids, but the numbering labels are inaccurate. Following the order from top to bottom and left to right, the transitions between adjacent stages show minor logical issues. A few adjacent stages are nearly repetitive, leading to stage redundancy."
"Final Score: 3"

Input:
- Image A: The first uploaded photo.
- Image B: The second uploaded photo.
- Instruction: {Instruction}

Output Format:
After completing the evaluation, please output the result in the following format(X is required to be an integer rating from 1 to 5):

Final Score: X

Figure 29. Prompt for evaluating Logical Coherence.

## Prompt for evaluating Scientific Plausibility

You are an image process evaluation expert with profound knowledge literacy, particularly skilled at accurately judging the rationality and correctness of process images based on real processes (such as underlying mechanisms, scientific principles, chemical reactions, key features, etc.). Please conduct a strict evaluation of the input Image B. You will receive the following inputs:
- Image A: Image A consists of two parts. On the left side (or top) is the reference start image, and on the right side (or bottom) is the reference end image.
- Image B: The "intermediate transition process" image generated based on the reference start and end images.
- Instruction: A description of how to transition from the reference start image to the reference end image to generate Image B, requiring Image B to fully reflect the intermediate process in grid format.
- Checklist: Compiled from scientific knowledge or key process features, listing point by point the details and elements that the intermediate process should cover.

Your task:
Evaluate, item by item, whether the content in Image B correctly expresses the key features listed in the checklist.

Scoring criteria (maximum score is 5):
To prevent lenient evaluations or default assumptions that the generated result is reasonable, please use strict standards to examine whether Image B has any deficiencies, omissions, or unclear expressions, and reflect these in your scoring. Do not assign high scores simply because of overall stylistic harmony or subjective speculation about intent. Scoring must be determined using the strictest and most conservative standards.
- 5 (Perfectly aligned): Image B perfectly presents all checklist items.
- 4 (Well aligned): Image B presents all checklist items well, with only minor deviations.
- 3 (Generally aligned): Image B presents all checklist items, though deviations exist, it still reasonably reflects the checklist.
- 2 (Largely misaligned): Image B does not present all checklist items, with missing elements and poor overall rationality.
- 1 (Completely misaligned): Image B fails entirely to meet the checklist requirements, losing overall rationality.

Evaluation guidance:
- If Image B merely replicates the start and end states provided in Image A without focusing on the intermediate process, then Image B does not meet the basic requirement of expressing the intermediate transition process.
- If Image B expresses the intermediate transition process, analyze the explicitly presented objective content of Image B based on the checklist and its descriptions, and evaluate how well Image B aligns with the checklist items.

Input:
- Image B: The first uploaded photo.
- Instruction: {Instruction}
- Checklist: {Checklist}

Output format:
After completing the evaluation, please output the result as follows(X is required to be an integer rating from 1 to 5):

Final Score: X

Figure 30. Prompt for evaluating Scientific Plausibility.

## Prompt for evaluating Process Plausibility

You are an image content analysis expert. Based on the following inputs, evaluate whether the model truly understands the "intermediate transition process from the reference start image to the reference end image." You will receive the following inputs:
- Instruction 1: Describe how to transition from the reference start image to the reference end image to generate Image B, including explicit intermediate transition path constraints.
- Instruction 2: Describe how to transition from the reference start image to the reference end image to generate Image C, including explicit intermediate transition path constraints.
- Image A: Composed of two parts—left/top as the reference start image, right/bottom as the reference end image.
- Image B: The intermediate transition process result generated from Image A's start/end images (should comply with the path constraints in Instruction 1).
- Image C: The intermediate transition process result generated from Image A's start/end images (should comply with the path constraints in Instruction 2).

Evaluation Task:
Determine whether the model truly understands and clearly expresses the intermediate transition process from start to end, strictly follows the path constraints in Instruction 1 and Instruction 2 respectively, and reflects differentiation between the two paths.

Scoring Criteria (Maximum 5 points):
Do not relax the standard due to overall stylistic harmony or subjective speculation of intent; score only based on explicitly presented content in Images B and C. Please do not assign a higher score simply because the overall style appears coordinated or reasonable. Use the strictest and most conservative standard for judgment.
- 5 points (Complete Understanding): Both B and C accurately, clearly, and with high quality reproduce the full transition process, strictly conforming to their respective path constraints; demonstrates strong understanding and differentiation ability.
- 4 points (Good Understanding): B and C reflect the transition process well, meet the corresponding path constraints, and show generally good understanding.
- 3 points (Average Understanding): B and C roughly present the transition process, basically reflect the path constraints, but contain inaccuracies.
- 2 points (Poor Understanding): B and C show transitions but lack clear path differentiation or fail to fully implement the constraints; unable to generate according to the required paths.
- 1 point (No Understanding): B and C cannot reasonably reflect the intermediate process, paths are invalid/chaotic, do not match the textual instructions.

Key Evaluation Points (Check item by item):
- Explicitness and completeness of intermediate process: (1) Do B and C clearly show "intermediate steps," rather than simply copying or slightly modifying the start/end states? (2) Steps must be presented sequentially in a grid format (each grid as one stage, with the stage number in the top-left corner); do not rely on common sense or assumed knowledge to fill in unexpressed steps.
- Conformance to path constraints (verify item by item): (1) In B and C, does each step explicitly correspond to the path constraints described in their respective instructions (explicit evidence only)? (2) "Looks reasonable overall" cannot substitute for explicit compliance.
- Path understanding and differentiation ability: (1) Under different path constraints, do B and C show distinct intermediate processes and stage sequences? (2) Check for skipped stages, redundant stages, or missing stages, and deduct points accordingly.

Examples:
- "B explicitly shows the intermediate process path, but deviates somewhat from the path requirements; C's final result fits, but intermediate steps contain stage skipping/redundancy, failing to reflect the complete path process."
- "Final score: 2"

Input:
- Instruction 1: {Instruction_A}
- Instruction 2: {Instruction_B}
- Image A: First uploaded photo.
- Image B: Second uploaded photo.
- Image C: Third uploaded photo.

Output format:
After completing the evaluation, please output the result in the following format:

Final score: X

Figure 31. Prompt for evaluating Process Plausibility.